AN INVESTIGATION OF PRE-SERVICE TEACHER ASSESSMENT LITERACY AND ASSESSMENT CONFIDENCE: MEASURE DEVELOPMENT AND EDTPA PERFORMANCE

A dissertation submitted to the
Kent State University College
of Education, Health, and Human Services
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

By

Kelli A. Ryan

Spring 2018

Graduate Program in Evaluation and Measurement

Kent State University
2018

A dissertation written by

Kelli A. Ryan

B.A., Ohio University, 2011

M.A., Kent State University, 2014

Approved by

____________________________, Director, Doctoral Dissertation Committee
Aryn C. Karpinski, Ph.D.

____________________________, Member, Doctoral Dissertation Committee
Erica Eckert, Ph.D.

____________________________, Member, Doctoral Dissertation Committee
Denise N. Morgan, Ph.D.

Accepted by

____________________________, Director, School of Foundations, Leadership, and
Kimberly Schimmel, Ph.D. Administration

____________________________, Dean, College of Education, Health and Human
James C. Hannon, Ph.D. Services

Kelli A. Ryan, May 2018 EVALUATION AND MEASUREMENT

AN INVESTIGATION OF PRE-SERVICE TEACHER ASSESSMENT LITERACY AND ASSESSMENT CONFIDENCE: MEASURE DEVELOPMENT AND EDTPA PERFORMANCE (319 pp.)

Director of Dissertation: Aryn C. Karpinski, Ph.D.

The need to create assessment literate and assessment confident teachers is increasing (Popham, 2009; 2011). Research has revealed that teachers are not well trained to use assessment in the classroom and are poorly trained in standardized testing (Zhang & Burry-Stock, 1997; Zhang & Burry-Stock, 2003). The purpose of this study was to: (1) evaluate the psychometric properties (i.e., reliability and validity) of an instrument that measures the assessment literacy and assessment confidence of pre-service teachers (i.e., the Classroom Assessment Literacy Inventory [CALI]), and (2) investigate the relationship between assessment literacy, assessment confidence, and scores on a performance-based assessment (edTPA).

In the pilot testing phase, Rasch Analysis and Rasch Principal Components Analysis (PCA) were used to evaluate the psychometric properties (i.e., reliability and validity) of the assessment literacy and confidence measures (i.e., the CALI). The pilot sample (*N* = 165) consisted of sophomores and juniors in one teacher preparation program in the Midwestern United States (US). After the pilot testing phase, the instrument was revised and administered to a second sample of 112 pre-service teachers who were in their final semester of the same undergraduate teacher preparation. Confirmatory Factor Analysis (CFA) was used to provide evidence of the internal structure of the CALI. Following the CFA, controlling for other demographic and academic variables such as teacher education program (e.g., Early Childhood, Middle Childhood, Adolescent Education, etc.) and Grade Point Average (GPA), the impact of the second phase sample's assessment confidence on the relationship between assessment literacy and performance-based assessment scores was examined.

Results indicated the limited range of the assessment-related content measured by the modified CALI, as well as the modified CALI's relative difficulty for this sample. Significant relationships were found between pre-service teacher Program and GPA on the relationship between assessment knowledge, assessment confidence, and a performance-based assessment. Discussion and implications for teacher education programs emphasizes the relationship between assessment knowledge and performance, GPA and performance, as well as the differences between programs on the main variables of interest. Methodological and statistical discussion and implications are presented for the use of Rasch PCA, parceling, the CFA model, and the benefits to considering a mixed-methods methodological approach.

ACKNOWLEDGEMENTS

To the incredible teachers and mentors I have had along the way, especially those in the Evaluation and Measurement program, this would not have been possible without your guidance and teaching. I am especially indebted to Dr. Aryn C. Karpinski, whose patience, passion, dedication, and empathy for her students has no bounds. Dr. Karpinski's work as a Dissertation director is unparalleled, and I am grateful for the privilege to work with her. Thank you to my committee members, Dr. Erica Eckert and Dr. Denise Morgan, for their feedback, as well as Dr. Jacob Barkley for serving as the Graduate Faculty Representative on my committee. I must also thank Dr. Phillip Hamrick for showing me that research methods and statistics are not so scary so afterall.

To my parents, thank you for your patience, love, and support. Being an eternal student might not have been what you envisioned for me, but I could not have done this without you. You made sure I had a roof over my head and then some. You were understanding of my academic career path despite its challenges.

Lastly, I must thank my soon-to-be husband, Parke, and my beloved dog, Jack. Parke, if you loved me during the Dissertation process, I have faith our marriage will last many years to come. Jack, you can't read this, but thank you for always being there when I really needed to pet something soft and see a happy face.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS iv

LIST OF FIGURES ix

LIST OF TABLES x

CHAPTER

I. INTRODUCTION 1

Purpose and Rationale 7

Research Questions 9

Implications 12

Summary 14

II. LITERATURE REVIEW 16

The Process of Assessment 17

Assessment Literacy 20

Definitions of Assessment Literacy 22

Existing Assessment Literacy Frameworks 26

Self-Efficacy 34

Summary 41

III. METHODOLOGY 43

Purpose 43

Context. 48

Participants. 51

Pilot sample 52

Second phase sample 53

Measures 54

The Classroom Assessment Literacy Inventory (CALI) 54

The Assessment Confidence Scale 62

The edTPA 63

Procedures. 69

Measure Development of the CALI 69

Measure Development of the Confidence Scale 70

Data Collection 71

The CALI 72

The Assessment Confidence Scale 74

The edTPA 75

Data Analysis. 76

Phase 1: Pilot Sample 76

Rasch analysis of the modified CALI 76

Rasch Principal Components Analysis (PCA) . . . 81
Phase 2: Confirmatory Study . . . 83
CALI Confirmatory Factor Analysis (CFA) . . . 83
Assumptions . . . 85
Estimation methods . . . 87
CFA sequence of steps . . . 88
Model specification . . . 89
Model identification . . . 89
Model estimation . . . 90
Model testing and modification . . . 91
CFA graphical and formulaic representation . . . 92
Multiple Regression . . . 92
Moderation Multiple Regression . . . 94
Moderated effect hypothesis . . . 96
Visualizing moderation . . . 98
Probing and interaction . . . 98
Summary . . . 99

IV. RESULTS. . . . 100
Research Question 1. . . . 101
Descriptives . . . 101
Rasch Analysis . . . 108
Pilot CALI Rasch analysis . . . 108
Pilot confidence Rasch analysis . . . 114
Measure refinement . . . 120
25-Item modified CALI results . . . 122
Rasch Principal Components Analysis (PCA) . . . 131
First component Rasch analysis . . . 135
First component confidence Rasch analysis . . . 139
Second component Rasch analysis . . . 146
Second component confidence Rasch analysis . . . 150
Component item characteristics summary. . . . 157
Research Question 2 . . . 163
Descriptives . . . 163
Confirmatory Factor Analysis (CFA) . . . 169
Model estimation . . . 169
Parceling . . . 170
Component 1 parcels . . . 172
Component 2 parcels . . . 172
Model identification . . . 173
Model modification . . . 176
Summary of All Psychometric Results . . . 179
Moderated Multiple Regression . . . 185
Model variables . . . 186

Outliers 191
Assumptions 192
Correlations 193
Moderated Multiple Regression Models 195
Overall assessment knowledge models – model 1 196
Overall assessment knowledge models – model 2 197
First component assessment knowledge models – model 3 198
First component assessment knowledge models – model 4 199
Second component assessment knowledge models – model 5 200
Second component assessment knowledge models – model 6 201
Summary 207

V. DISCUSSION 208
Research Question 1 and 1A 210
Demographic and Descriptive Information (Pilot Sample) 211
Assessment Literacy Measure Development (Pilot Sample) 213
Unidimensional assessment literacy: comparing the 35-Item and 25-Item measures 214
Unidimensional assessment confidence: comparing the 35-Item and 25-Item measures 215
Unidimensional assessment confidence: comparing the 35-Item and 25-Item measures 217
Multidimensional assessment literacy: comparing the two-component and 25-item unidimensional measures 218
Multidimensional assessment confidence: comparing the two-component and 25-item unidimensional measures 220
Research Question 2 222
Demographic and Descriptive Information (Second Phase Sample) 222
Assessment Literacy Measure Internal Structure Confirmation (Second Phase Sample) 225
Internal structure of the CALI with Confirmatory Factor Analysis (CFA) 225
Moderating the Assessment Literacy and Performance Assessment Relationship 227
Moderated Multiple Regression models 1 through 6 227
Exploring no moderation and significant covariates 229
Implications 232
Teacher Preparation Programs 232
Psychometricians and Researchers 242
Limitations 245
Conceptual Limitations and Future Directions 245
Methodological Limitations and Future Directions 247
Statistical and Psychometric Limitations and Future Directions 251

Conclusion 255

APPENDICES 258
APPENDIX A. MODIFIED CLASSROOM ASSESSMENT LITERACY INVENTORY (CALI) 259
APPENDIX B.CLASSROOM ASSESSMENT LITERACY INVENTORY (CALI) ITEMS AND THE STANDARDS 277
APPENDIX C. INSTITUTIONAL REVIEW BOARD (IRB) FORMS 279
APPENDIX D. ITEM RE-NUMBERING AND PRINCIPAL COMPONENTS ANALYSIS (PCA) COMPONENTS 290
APPENDIX E. ITEMS ACCORDING TO PRINCIPAL COMPONENTS ANALYSIS (PCA) COMPONENT LOADINGS 292

REFERENCES 297

LIST OF FIGURES

Figure Page

1 edTPA Assessment Rubric Example – Rubric 15 66

2 Confidence as a Moderator of Assessment Knowledge and Performance Outcomes 97

3 Pilot Sample 35-Item Item-Person Map 110

4 Pilot Sample 35-Item Confidence Item-Person Map 114

5 Pilot Sample 35-Item Confidence Category Probabilities 119

6 Pilot Sample 25-Item Item-Person Map 123

7 Pilot Sample 25-Item Confidence Item-Person Map 127

8 Pilot Sample 25-Item Confidence Category Probabilities 130

9 PCA Figure with Variance and Components 132

10 Pilot Sample 12-Item CALI Component 1 Item-Person Map 136

11 Pilot Sample 12-Item CALI Component 1 Confidence Item-Person Map 140

12 Pilot Sample 12-Item Component 1 Confidence Category Probabilities 145

13 Pilot Sample 13-Item Component *2* Item-Person Map 147

14 Pilot Sample 13-Item Component 2 Confidence Item-Person Map 151

15 Pilot Sample 13-Item Component 2 Confidence Category Probabilities 156

16 CFA Graphical Representation 174

17 CFA Final Model Graphical Representation 178

LIST OF TABLES

Table Page

1 Popham’s (2011) Suggested Content for Teacher Assessment Literacy 22

2 Pilot Sample Variable Descriptive Statistics: Classroom Assessment Literacy Inventory (CALI) Total Scores and Assessment Confidence Total (Average) Scores (N = 165) 105

3 Relationships between Pilot Sample Variables and Classroom Assessment Literacy Inventory (CALI) Total Scores and Assessment Confidence Total (Average) Scores (N = 165). 107

4 Pilot Sample 35-Item CALI Summary of Person Statistics (N = 165) 111

5 Pilot Sample 35-Item CALI Summary of Item Statistics (N = 35) 111

6 Pilot Sample 35-Item CALI Item Misfit Statistics (N = 35) 112

7 Pilot Sample 35-Item Confidence Measure Summary of Person Statistics (N = 165) 116

8 Pilot Sample 35-Item Confidence Measure Summary of Item Statistics (N = 35) . . 116

9 Pilot Sample 35-Item Confidence Measure Item Misfit Statistics (N = 35) 117

10 Pilot Sample 35-Item Confidence Measure Category Thresholds and Fit Table (N = 35) 119

11 Pilot Sample 25-Item CALI Summary of Person Statistics (N = 165) 124

12 Pilot Sample 25-Item CALI Summary of Item Statistics (N = 25) 124

13 Pilot Sample 25-Item CALI Misfit Statistics (N = 25) 125

14 Pilot Sample 25-Item Confidence Measure Summary of Person Statistics (N = 128

165)

15 Pilot Sample 25-Item Confidence Measure Summary of Item Statistics (N = 25) 128

16 Pilot Sample 25-Item Confidence Measure Misfit Statistics (N = 25) 129

17 Pilot Sample 25-Item Confidence Measure Category Thresholds and Fit Table (N = 25) 131

18 Pilot Sample Rasch Principal Components Analysis (PCA) CALI Item Loadings (N = 25) 134

19 Pilot Sample 12-Item CALI Component 1 Summary of Person Statistics (N = 165) 137

20 Pilot Sample 12-Item CALI Component 1 Summary of Item Statistics (N = 25) 138

21 Pilot Sample 12-Item CALI Component 1 Item Misfit Statistics (N = 12) 139

22 Pilot Sample 12-Item Component 1 Confidence Measure Person Statistic (N = 165) 142

23 Pilot Sample 12-Item Component 1 Confidence Measure Item Statistic (N = 12) 142

24 Pilot Sample 12-Item Component 1 Confidence Measure Misfit Statistics (N = 12) 143

25 Pilot Sample 12-Item Component 1 Confidence Measure Thresholds and Fit Table (N = 12) 145

26 Pilot Sample 13-Item CALI Component 2 Summary of Person Statistics (N = 165) 148

27 Pilot Sample 13-Item CALI Component 2 Summary of Item Statistics (N = 13) 149

28 Pilot Sample 13-Item CALI Component 2 Item Misfit Statistics (N = 13) 150

29 Pilot Sample 13-Item Component 2 Confidence Measure Summary of Person Statistics (N = 165) 153

30 Pilot Sample 13-Item Component 2 Confidence Measure Summary of Item

Statistics (N = 13) 153

31 Pilot Sample 13-Item Component 2 Confidence Measure Misfit Statistics (N = 13) 154

32 Pilot Sample 13-Item Component 2 Confidence Measure Thresholds and Fit Table (N = 13) 156

33 Component Item Characteristics Summary 158

34 Second Phase Sample Variable Descriptive Statistics: Classroom Assessment Literacy Inventory (CALI) Total (25 Item) Scores and Assessment Confidence (Average) (25 Item) Scores (N = 112) 166

35 Relationships between Second Phase Sample Variables and Classroom Assessment Literacy Inventory (CALI) Total Scores (25 Item) and Assessment Confidence Total Scores (N = 112) 168

36 Summary of Parcel Information (N = 8). 173

37 Summary of Factor Loading Information (N = 8) 176

38 Model Modification and Fit Index Comparison Information 178

39 Component 1 Knowledge and Confidence: Second Phase Sample Variable Descriptive Statistics for (12 Item) Classroom Assessment Literacy Inventory (CALI) Scores and Assessment Confidence (Average) Scores (N = 112) 180

40 Component 2 Knowledge and Confidence: Second Phase Sample Variable Descriptive Statistics for (13 Item) Classroom Assessment Literacy Inventory (CALI) Scores and Assessment Confidence (Average) Scores (N = 112) 182

41 Second Phase Sample Variables and Classroom Assessment Literacy Inventory (CALI) Component 1 Score and Confidence Component 1 Score (12 Item)

Relationships (N = 112) 184

42 Second Phase Sample Variables and Classroom Assessment Literacy Inventory (CALI) Component 2 Score and Confidence Component 2 Score (13 Item) Relationships (N = 112) 184

43 Main Moderated Multiple Regression Models 185

44 Second Phase Sample Variable Descriptive Statistics: edTPA Overall Scores and edTPA Assessment Domain Scores (N = 112) 189

45 Relationships between Second Phase Sample Variables and edTPA Overall Scores and edTPA Assessment Domain Scores (N = 112) 191

46 Correlation Matrix for Second Phase Outcome Variables (N = 112) 194

47 Moderated Multiple Regressions Models 1 and 2: Predicting edTPA Total and Assessment Scores from CALI Total Scores Moderated by Assessment Confidence Scores (N = 112) 203

48 Moderated Multiple Regressions Models 3 and 4: Predicting edTPA Total and Assessment Scores from Component 1 CALI Scores Moderated by Component 1 Assessment Confidence Scores (N = 112) 204

49 Moderated Multiple Regressions Models 5 and 6: Predicting edTPA Total and Assessment Scores from Component 2 CALI Scores Moderated by Component 2 Assessment Confidence Scores (N = 112) 205

50 Summary of Key Research Findings 231

51 Summary of Program Performance on Three Outcome Variables 238

CHAPTER I

INTRODUCTION

Assessment literacy, as defined by Popham (2011), is an individual's understanding of the fundamental assessment concepts and procedures deemed likely to influence educational decisions both in the classroom (i.e., classroom assessment) and those that impact the inside and outside of the classroom (i.e., accountability assessment). Research has shown that teachers spend about half of their time involved in assessment-related activities, highlighting the need to prepare assessment literate educators (Plake, Impara, & Fager, 1993; Stiggins, 1991). Because of the amount of classroom time devoted to these practices, teacher competency in measurement and assessment is essential to the success of not only the teacher, but also the students (Zhang & Burry-Stock, 2003). However, previous reviews of literature on assessment knowledge measures has shown that teachers are not well trained to use assessment in the classroom (e.g., standardized tests), with the majority of teachers engaging in inappropriate practices of teaching test items, increasing time limits, giving hints, and changing students' answers (Zhang & Burry-Stock, 1997). Moreover, previous research indicates that K-12 teachers use a range of assessment methods in the classroom (i.e., formal and informal), but are poorly trained in the administration and interpretation of standardized tests, a measurement-related area of assessment (Zhang & Burry-Stock, 2003).

Assessment literacy taskforces such as the Michigan Assessment Consortium (MAC) – a well-known professional association of educators focused on the use of accurate, balanced, and meaningful assessment – aim to create educator understanding of the use and practice of assessment. The MAC defines assessment literacy as including, but not limited to, tasks such as communicating and understanding assessment results, selecting, creating, and evaluating assessments, and assessment-related decision making (Michigan Assessment Consortium [MAC], 2015). Popham's (2011) definition of assessment literacy, which is the focus of this study, considers these skills, but within the context of the individual's understanding of the fundamental assessment concepts and procedures deemed likely to influence educational decisions both in the classroom (i.e., classroom assessment) and those that impact the inside and outside of the classroom (i.e., accountability assessment). Connecting the importance of assessment literacy with a teacher's numerous roles involving assessment addressed above, focusing on assessment in teacher preparation is indispensable (Popham, 2009).

Some measures of assessment literacy for a variety of educator groups currently exist and have provided evidence of the many concepts subsumed under assessment literacy. One measure is the Classroom Assessment Literacy Inventory (CALI; Mertler, 2003), which consists of 35 multiple-choice questions aligned with the seven *Standards for Teacher Competence in Educational Assessment of Students* developed by the American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), and the National Education Association (NEA; 1990). Specifically, the seven *Standards* outline what it means to be an assessment-literate teacher, stressing competence in: (1) choosing assessment methods, (2) developing assessment methods, (3) administering, scoring, and interpreting assessment results, (4) using assessment results for decision making, (5) grading, (6) communicating assessment results, and (7) recognizing unethical practices (AFT, NCME, & NEA, 1990; Zhang, 1996). The *Standards*, although not the only barometer of teacher assessment competency, describe the extent to which a teacher is assessment literate (Stiggins, 1999). The CALI, and other similar measures based on the *Standards*, was developed to provide evidence of assessment literacy and facilitates measurement of this construct in pre-service teachers. Furthermore, the *Standards* share considerable overlap with definitions of assessment literacy such as Popham's (2011), providing evidence of alignment between the construct of assessment literacy and existing measures of assessment literacy, such as the CALI.

Arguably, to apply assessment knowledge effectively, a teacher must also be confident in their understanding of assessment. Sociological and psychological theories, such as those developed by Bandura (1977), outline the relationship between confidence and/or self-efficacy and performance. Bandura's Social Cognitive Theory suggests that when an individual feels competent and has confidence in completing a task, he/she will choose to engage in it. On the other hand, when an individual feels incompetent and lacks confident in a task, he/she will avoid engaging in it.

Applying Badura's (1977) definition to educators, it is vital that teachers feel competent (i.e., assessment literate) and confident (i.e., assessment confidence) in their assessment abilities in order to engage in the process. Several factors such as practice, exposure, and application contribute to how confident an individual is when using or applying a skill to a new context, according to the Social Cognitive Theory (Bandura, 1977). For teachers, this means that when he/she has a solid understanding of a concept such as communicating assessment results to parents, he/she is more likely to be successful in their use and application of assessment. Therefore, to create assessment literate educators, assessment confidence must be addressed. This is of particular importance with new teachers who are just beginning to apply the knowledge they learned from student teaching and coursework to their own classrooms in the first few years after completing a teacher preparation program.

Existing educational research has focused largely on how confident teachers are in their general training, content knowledge, pedagogical content knowledge, and areas of practice like classroom management (e.g., Burgoyne, Cantrell, Smith, & St. Garbett, 2003; Bursal & Paznokas, 2006; Clair, & Harris, 2009; Dembo & Gibson, 1985; Graham & Watson, 2001). The majority of teacher self-efficacy studies focus on specific content areas like math, science, technology, and foreign language teaching. This focus on explicit content areas reflects the large belief that a teacher must possess high self-efficacy, specifically in their exclusive content domain knowledge. Such research studies suggest that teacher confidence is relative to specific content areas within their teaching specialty (Bursal & Paznokas, 2006; Garbett, 2003). Other areas of pedagogy such as classroom management have yielded similar results (Main & Hammond, 2008). However, teacher confidence and knowledge specific to assessment abilities has yet to be explored. It is possible that assessment is both embedded in how teachers understand the content they are teaching (i.e., content knowledge) and how they teach (i.e., pedagogical content knowledge). Therefore, investigation into the presence and impact of assessment confidence is warranted.

Research on confidence and performance also indicates that there is a relationship between self-efficacy and competence (e.g., Bursal & Paznokas, 2006; Garbett, 2003; Watson, 2001; Wolters & Daugherty, 2007). That is, confident teachers generally have higher levels of competency in their specific content area specializations. The connection between pre-service teachers' assessment literacy and assessment confidence and its impact on their performance (e.g., on portfolio-based assessments) has thus far been understudied. As the use and presence of assessment continues to increase, so does its impact on teachers, both in teacher preparation programs as well as the K-12 classroom. Examining the relationship between assessment literacy and confidence in relation to performance can provide research-based evidence to assist in teacher preparation.

The need to create assessment literate and assessment confident teachers is increasing (Popham, 2009; 2011). Institutions of higher education are at the forefront of this change as they experience shifts in state policy towards performance-based exams like the edTPA, which measures pre-service teacher readiness for teaching using a series of fifteen rubrics across three core domains: (1) Planning, (2) Instruction, and (3) Assessment. This performance-based portfolio assessment of teacher preparedness is submitted by pre-service teachers across approximately 741 higher education institutions or state entities, including 39 states and the District of Columbia, as of Spring 2017. In these states, the edTPA exam is required in some capacity, ranging from initiating implementation to current state-wide licensure requirements, for matriculating pre-service teachers (edTPA.org). As many states take steps toward total statewide implementation, undergraduate teacher education programs are beginning to evaluate the use of the edTPA exam and its components relative to pre-service teacher preparation.

While no two teacher education programs are identical, they include methodological and pedagogical courses which teach material on how to teach specific content and subjects like reading, math, and science. The edTPA not only assesses student understanding and ability related to one content area (Stanford Center for Assessment, Learning and Equity [SCALE], 2013), it also evaluates their understanding of classroom assessment, as it is a core domain of the exam. Focusing on assessment knowledge and/or literacy emphasizes the significance of this topic in the national discussion of teacher preparation (SCALE, 2017). This exam is performance-based and therefore requires students to submit a portfolio of information including lesson plans, reflections on classroom performance, and videos of classroom exercises and lessons. The requirements for what constitutes a passing or failing score vary by state or institution. All available state scoring information is freely available; however, it only exists for states that have implemented edTPA at the state level (i.e., see edTPA.com for all scoring determinations).

The increased importance of assessment knowledge and confidence in not isolated to presence of the edTPA alone. Teachers are inundated with assessment-related activities at the classroom, district, state, and national level. Teachers are not only responsible for creating, conducting, measuring, and evaluating their own classroom assessments, but they are also accountable for preparing their students for district, state, and national assessments (Popham, 2003; 2011). Additionally, teachers are in charge of administering these exams and explaining the purposes, uses, and results to their students, the parents of their students, and the community. Given the current and increasing assessment-related demands on teachers, evidence from a study on assessment knowledge and the role of assessment confidence in relation to a performance-based assessment could provide teacher education programs with best practices for pre-service teacher education.

Purpose and Rationale

The first objective of this study is to investigate the psychometric properties (i.e., reliability and validity) of the modified CALI, from the original CALI developed by Mertler (2003), using a sample of undergraduate students in a teacher education program at a large Midwestern university in the U.S. (i.e., Ohio). Specifically, this study will examine assessment knowledge and understanding (i.e., assessment literacy) within a sample of pre-service teachers by investigating the psychometric properties (i.e., (i.e., reliability and validity)) of participant scores on the CALI. Additionally, this study includes a measure of confidence after each question on the CALI (i.e., assessment confidence). Thus, the proposed study will evaluate assessment literacy in pre-service teachers and how confident he/she is in their assessment knowledge.

The second objective of this study is to investigate the impact of assessment confidence on the relationship between assessment literacy and performance assessment scores. All pre-service teacher education students within the target population were required to take the edTPA performance assessment as a graduation requirement. The edTPA consists of several sub-scales and fifteen rubrics designed to evaluate pre-service teacher readiness across the domains of Planning, Instruction, and Assessment. At this time, the edTPA is not a requirement for licensure in the state of Ohio. However, due to the increased national use of the edTPA exam, the proposed study will evaluate the relationship between classroom assessment knowledge, assessment confidence, and high-stakes performance assessment outcomes.

From the abovementioned two objectives, the present study will investigate assessment literacy and assessment confidence within a sample of pre-service teachers at a large state university in the Midwest. Participants in this study were at the point of graduation, which provides insight into pre-service teachers who are nearing the end of their undergraduate formal education. That is, a sample of students approaching the transition from pre-service to in-service teaching was recruited, which included those in the final months of their teacher preparation program and those who submitted their edTPA portfolio assessment. This sample and the two objectives provide an evaluation of the level of assessment knowledge in the average emerging in-service teacher, not only in relation to performance outcomes (i.e., edTPA), but also based on an external, objective assessment knowledge measure (i.e., the CALI). Research into the usefulness of the CALI also provides insight into the development of measures of assessment literacy and assessment confidence, which offers evidence for implementing changes to prepare new teachers.

As noted above, the rationale for the present study stems from the increased presence of assessment in K-12 classrooms (and beyond) and assessment's impact on teacher education. Specifically, higher education institutions, such as the target population in this study, are adjusting to changes made at the state and national level that impact teacher preparation. The edTPA is one example of this state-level change that is impacting teacher education programs. These programs aim to prepare successful teachers, which states define as those who pass licensure requirements.

Research on current pre-service teacher assessment knowledge is necessary in order to investigate the hypothesized relationship between assessment knowledge and the edTPA's measurement of performance-based assessment knowledge as a high-stakes licensure exam. The possible impact of confidence must also be explored in order to consider contributing factors of performance. In addition, results from studies on the relationship between assessment knowledge and performance-based assessment can be used to evaluate teacher preparation courses and programs that use student performance as an opportunity for curricular change. Thus, exploring the relationship between confidence and assessment knowledge is important because it can promote the preparation of successful pre-service teachers who pass new licensure exams, such as the edTPA.

Research Questions

This study has two main research questions. The first research question aligns with the first research objective examining the psychometric properties (i.e., reliability and validity) of the CALI. Research Question 1 (RQ1) states, "What are the psychometric properties of the newly-developed assessment literacy and confidence measure for pre-service teachers?" In this study, the modified CALI is an adjusted version of the 35-question multiple-choice assessment created by Mertler (2003; Mertler & Campbell, 2005). The modified CALI includes a confidence rating after each of the multiple-choice questions. The confidence rating asks participants to rate how confident he or she is in their response (i.e., assessment knowledge) on a 5-point Likert scale. The modification of this measure and its development will be discussed in detail in Chapter 3. Rasch Analysis was used to analyze the responses to the assessment knowledge and confidence items.

Additionally, through the CALI's alignment with the seven *Standards for Teacher Competence in Educational Assessment of Students*, a related and subsequent research question (RQ1A) asked: "What is the internal structure of the modified CALI?" Even though the measure is based on the seven areas of teacher assessment knowledge, as determined by the organizations and individuals responsible for creating the *Standards*, this does not imply that there are seven components, dimensions, or factors of knowledge comprising the construct. Rasch Analysis was used to determine the internal structure of the items via a Principal Components Analysis (Rasch PCA) of the residuals. The CALI items were analyzed separately for both content questions and confidence questions. Assessment knowledge and confidence components were then investigated further in the second research question.

The second research question (i.e., related to the second research objective) examines the impact of assessment literacy and assessment confidence on performance assessment scores. Specifically, Research Question 2 (RQ2) stated, "What is the impact of assessment confidence on the relationship between pre-service teachers' assessment literacy and performance assessment scores?" Assessment confidence is operationalized as pre-service teachers' confidence in their answers on the CALI, which measures classroom assessment literacy across several components. This study investigated how self-ratings of assessment confidence impact scores on an assessment knowledge measure (i.e., the CALI) and scores measuring assessment "performance" on a portfolio-based assessment (i.e., the edTPA).

Based on the Rasch PCA, additional Confirmatory Factor Analyses (CFAs) were conducted on the assessment knowledge data. Following an investigation of the internal structure, RQ2 was analyzed using Moderated Multiple Regression Analyses. A moderator variable is one that influences the strength of a relationship between two other variables (Keith, 2009). In the current study, confidence is a hypothesized moderator of assessment literacy as measured by the CALI and performance assessment outcomes related to assessment as measured by the edTPA.

Overall, assessment literacy is hypothesized to be important to all areas of teacher performance (i.e., planning, instruction, and assessment), but most notably to assessment practices. In addition, assessment knowledge is hypothesized to be foundational for the knowledge and skills that pre-service teachers need to succeed in other areas of the profession. These considerations regarding assessment literacy and teachers are consistent with views outlining the impact of teacher confidence, content knowledge, and pedagogical knowledge on teacher assessment literacy. Thus, there is research to support the hypothesized direct relationship between assessment literacy and edTPA outcomes. However, assessment confidence is the posited moderator between assessment literacy and edTPA performance. The hypothesized relationship between assessment confidence and assessment literacy is based on Bandura's (1977) Social Cognitive Theory, which provides the rationale that experience and exposure to any subject matter influences confidence. Furthermore, the idea that knowledge influences confidence stems from Bandura (1977), who hypothesized that confidence requires competence (i.e., knowledge) and that a relationship between confidence and knowledge exists.

Implications

This study's target population includes students in their last semester of an undergraduate teacher education program, which typically consists of final coursework, student teaching (i.e., direct experience whereby the student is assisting with or leading a classroom in K-12 approximately full-time), and licensure requirements. This group of students, at the culmination of their undergraduate teacher preparation, is representative of what pre-service teachers learn about assessment in their undergraduate program, how confident they are in this knowledge, and how they perform on external measures of teaching ability.

Results from this study provide feedback to teacher preparation programs on student preparation involving assessment (i.e., before entering the classroom). By evaluating their understanding of assessment, results from this study can assist universities in the review of their curricula, and how their teacher preparation courses and program have impacted student preparedness specifically related to assessment. Findings from this study may serve as a model for other institutions of higher education preparing K-12 teachers and adapting to edTPA changes. This model presents a way for teacher educators to identify areas of weakness in assessment knowledge, as well as the relationship between students' confidence and knowledge in engaging with the assessment process. This information may lead to curricular changes, offering assessment workshops, or embedding hands-on assessment experiences in courses to increase confidence.

Confidence (in general), self-efficacy, and self-concept have been investigated in previous teacher education studies, with no attention given specifically to assessment confidence (e.g., Brookhart, Loadman, & Miller, 1994; Bursal & Paznokas, 2006; Demo & Gibson, 1985; Garbett, 2003; Walters & Daughter 2007). Some research on the construct of assessment confidence exists but was limited to teacher confidence in the assessment of student learning, not in teachers' overall confidence in their assessment abilities (Leahy, Lyon, Thompson, & William, 2005). The majority of existing confidence research has largely explored pedagogical content domains such as math and science. That is, specific teaching content areas/specializations have been studied in isolation, and therefore only confidence related to that specific area or subject matter has been reported (Volante & Fazio, 2007).

The sample in the current study includes pre-service teachers across a variety of teacher preparation program content areas (e.g., Early Childhood, Middle Childhood, and Adolescent Science Education, Math Education, etc.). This composition allows for an investigation of assessment confidence across different content-domains. Therefore, this study's exploration of the relationship between assessment confidence and assessment knowledge considers possible content area or program level differences. This information provides teacher education programs with a foundation for preparing assessment-ready and confident teachers. It also provides evidence for the construct of assessment confidence and how it relates to knowledge and program differences.

Lastly, this study also examined assessment literacy and assessment confidence in relation to performance-based portfolio assessment scores (edTPA). The use of large-scale, standardized performance-based assessment of pre-service teachers is increasing, as shown by the national growth of edTPA, and the development of state-level performance assessments, like the Performance Assessment for California Teachers (Newton, 2010). In particular, the edTPA performance assessment is currently being used in some capacity in over 30 states. Studying the relationship between edTPA performance, which includes a specific assessment section, along with pre-service teacher knowledge of and confidence in assessment can provide some insight into how well teacher preparation programs are preparing their students. This becomes increasingly important as the adoption of edTPA continues nationwide.

Summary

The purpose of this chapter was to present a general overview of the background, purposes, rationale, and contributions of this study. It also presented the definition of assessment literacy used in this study. Assessment literacy, as defined by Popham (2011), is an individual's understandings of the fundamental assessment concepts and procedures deemed likely to influence educational decisions inside and outside the classroom. The importance of research in the area of pre-service teacher preparation, specifically related to assessment, was emphasized. As teaching standards at the national level evolve, higher education institutions are adjusting to such changes by including other requirements such as portfolio-based assessments of pre-service teacher performance (i.e., the edTPA). The following chapter (i.e., Chapter 2: Literature Review) will present the relevant existing research on the two main areas from this overview – assessment literacy and confidence.

CHAPTER II

LITERATURE REVIEW

Assessment is any range of methods used for the purposes of evaluating learner performance or attainment, including but not limited to formative assessment, summative assessment, and standardized assessment (Popham, 2013). Additionally, educational assessment serves the function of supporting learning (Wiliam & Leahy, 2007). Successful teachers use assessment to create meaningful information to be used as feedback, inform instruction, and modify pedagogical practice to benefit their students (Angelo & Cross, 1991; Black & Wiliam, 1998). Under these definitions, assessment comprises various activities such as classroom observation, discussion, and evaluating student work including in-class assignments, homework, and exams.

The relationship between teachers and assessment is one of the most important roles of a teacher in the classroom. Despite the importance of assessment in the classroom, the majority of teachers do not feel adequately prepared to engage in assessment and assessing student performance (Darling-Hammond, Chung, & Frelow, 2002; Mertler, 1998; 1999; Mertler & Campbell, 2005; Stiggins, 1999). Feelings of discomfort associated with assessment and the lack of assessment preparation across K-12 teachers were reviewed in Popham's (2003) study, which detailed the limited amount of teachers' assessment knowledge, or assessment literacy. Confusion, discomfort, and lack of assessment knowledge have been reported in pre-service teacher populations (Kahl, Hofman, & Bryant, 2013). However, these issues also appear in in-service teacher responses to assessment surveys, which report that 85% of K-12 teachers, of which nearly half were Adolescent Education teachers, do not feel prepared to assess student learning (Mertler, 1999).

More recently, this sentiment was echoed in an annual review by the National Council on Teacher Quality (NCTQ). Teachers' understanding and application of assessment was reported as lacking, with reports of less than 25% of K-12 teachers being prepared to use assessment in the classroom (NCTQ, 2013). Given these data and the importance of assessment in teacher education and practice, it is necessary to review several assessment-related concepts and processes. The following review will outline the process of assessment, assessment literacy, and several components that teachers need in order to be prepared to assess student learning.

The Process of Assessment

The assessment process involves the collection and evaluation of sufficient evidence needed to answer a specific question (DePascale, Betebenner, Ryan, & Sharp, 2017). Assessment knowledge is, in part, understanding how to execute this process. The assessment process can be ordered into these major steps: (1) Asking a question or determining a decision to be made, (2) Gathering information, (3) Determining if the information is sufficient to answer the proposed question(s), and (4) Using the evidence to answer the question or make the decision (DePascale et al., 2017; Suskie, 2009). From these steps above, the process of assessment can be viewed as circular in nature.

When an educator gathers information and evidence to answer a question, a critical point in the assessment process is determining when sufficient evidence has been collected to answer that question. Educators determine when sufficiency has been reached and when additional information needs to be collected. Thus, the knowledge and skills to execute the assessment process are incomplete without the ability to determine when sufficient evidence has been obtained, which is “the cornerstone of assessment literacy” (DePascale et al., 2017).

“Sufficiency” governs the presence and quality of evidence needed to substantiate a claim, make a decision, or answer a question (DePascale et al., 2017). Additionally, sufficiency, in this context, follows the adage “quality over quantity” in that evidential appropriateness is superior and the amount of evidence is ancillary (DePascale et al.). Educators must then be able to determine when sufficient evidence has been collected, which can be a difficult to determine as it is left to their discretion. However, teachers can use their contextual knowledge (i.e., the assessment, purposes, and outcomes) to define sufficiency in order to advance the assessment process or cycle.

A comparable process of assessment was introduced by Wiliam and Black (1996), who define assessment as encompassing three main parts: (1) Evidence, (2) Interpretation, and (3) Action. While Wiliam and Black (1996) share similar views on evidence as DePascale and colleagues (2017), the cycles are not equivalent. The first component of Wiliam and Black’s (1996) cycle states that the general level of an individual’s performance must be measured prior to creating inferences or actions. Assessing performance requires evidence, which can typically encompass artifacts such as writing samples, tests and quizzes, or audio or videotapes. In contrast, DePascale and colleagues (2017) specify that evidence collection is successive to question definition.

The second part of Wiliam and Black's (1996) assessment cycle is interpretation of the evidence or performance. Generally, student evidence is evaluated by the teacher, who is typically required to interpret the results in comparison to the student's classroom, grade level, school, district, state, or national standards. Therefore, in the context of the teacher and the classroom, the interpretation component of Wiliam and Black's (1996) assessment process is analogous to DePascale and colleagues' description of "sufficiency determination." Interpretation and the determination of sufficiency both involve teacher appraisal of the quality, quantity, and breadth of evidence needed to progress to action, or if additional performance assessment and data collection are warranted. These steps of the process lead to two possible conclusions: (1) Completion of the assessment process, or (2) Collection of more evidence. That is, if a teacher decides that he/she has enough information and evidence to answer a question (i.e., if his/her students have met the learning objective), then the process is complete. If the teacher decides that he/she has not collected enough information or the types of information necessary, he/she returns to collect further evidence.

Both processes of assessment outlined above by Wiliam and Black (1996) and DePascale and colleagues (2017) identify the third step of the assessment process as the decision or action phase. The final phase is where the evidence collected is used to answer a question or initiate change, ranging from pedagogy to policy. This last stage can be applied and adapted to a variety of contexts depending on the role of the educator. For teachers, this phase is not the same at it is for administrators or policy makers. The first conclusion is for the teacher to simply answer a question using the evidence that he/she has collected. An example of this occurs when a teacher determines if a student should or should not be a part of an accelerated program (e.g., advanced algebra). He/she assesses the student's capabilities and makes a conclusion based on evidence of the student's abilities. On the other hand, the teacher may use certain information to institute a change. This path of the assessment process engages the teacher in formative assessment, where he/she is using information to inform his/her practice.

Other similar notions and models of assessment exist (e.g., Reynolds, Livingston, Wilson, & Wilson, 2010; Suskie, 2009; Taras, 2005). The above brief overview of assessment and the process of assessment has centered on teachers' understanding of assessment. The understanding of assessment is often called "assessment literacy." Educators each have their own level of assessment literacy based on their roles. For teachers, assessment literacy is the understanding of basic assessment-related concepts that teachers need to be successful inside and outside the classroom. These concepts can range from understanding and conducting formative and summative assessments to communicating standardized test score reports. The following section will provide an overview of assessment literacy and the skills and competencies that teachers often need to be assessment literate.

Assessment Literacy

According to the National Council on Teacher Quality (2013), teacher education preparation programs are deficient in equipping future educators with assessment skills (Greenberg, McKee, & Walsh, 2013). In the Council's review of 690 teacher education program syllabi, only 24% were noted to adequately train teachers in how to assess learning and use student performance data to inform instruction. Thus, the majority of pre-service teachers may not have opportunities to fully use data derived from assessments or understand how to use these data to plan instruction. In a report on teacher preparation programs, Greenberg, McKee, and Walsh (2013) recommended that pre-service teachers be provided with "…multiple and rich course material in their preparation that will enable them to become assessment-literate and data-wise" (p. 21). Therefore, the call from these authors for increased quantity and quality of course material, compounded by the Council's report of overall preparation program deficiencies, magnifies the chasm between knowledge and application.

Recently, the influx of assessment results has driven teacher education and professional development programs to support and better prepare teachers to manage data. Gerzon (2015) advocated for administrators to encourage in-service teachers' assessment competencies via conducting applied data analysis. However, few studies exist describing examples of successful teacher education and professional development programs that have incorporated data management and applied analysis (e.g., Horn & Little, 2010; Lachat & Smith, 2005; Love, 2004; Supovitz & Klein, 2003; Suskie, 2009). Naturally, teachers often turn to their personal philosophies and their strong foundation of prior knowledge in decision making, creating a new challenge for brining measurement and data concepts into teacher preparation programs (Coburn, Honig, & Steign, 2009; McMillian, 2003). Teacher training programs also face logistical and practical restrictions that make instituting applied analysis difficult in most higher education settings. The following section will consider this context and present the definitions of assessment literacy detailed in the literature.

Definitions of Assessment Literacy

To date, no consensus on the definition of assessment literacy exists. Most definitions refer to literacy contextually or within a specific content area. A comprehensive overview of researchers' definitions of assessment literacy can be found in Xu and Brown (2016). The most widely used method in defining assessment literacy consists of listing specific assessment-related knowledge, understanding, and skills that an assessment literate educator must possess (e.g., Boyles 2005; Gareis & Grant, 2015; Popham, 2004; Stiggins, 1991; 2002; Xu & Brown, 2016).

One widely-cited and popular definition of assessment literacy (i.e., list) was developed by Popham (2011). Popham expanded an existing version of ideas on assessment literacy for educators developed by the Michigan Assessment Consortium (MAC), which is a taskforce of professionals in the field. Popham's (2011) list was created as a suggestion of what content teachers need to be assessment literate. These thirteen criteria are provided in Table 1 (Popham, 2011, p. 8-10). These points outline what Popham (2011) believed would comprise of thorough training and professional development.

Table 1

Popham's (2011) Suggested Content for Teacher Assessment Literacy

Criteria	Explanation
1	The fundamental function of educational assessment, namely, the collection of

evidence from which inferences can be made about students' skills, knowledge, and affect

2 Reliability of educational assessments, especially the three forms in which consistency evidence is reported for groups of test-takers (stability, alternate-form, and internal consistency) and how to gauge consistency of assessment for individual test-takers

3 The prominent role three types of validity evidence should play in the building of arguments to support the accuracy of test-based interpretations about students, namely, content-related, criterion related, and construct-related evidence

4 How to identify and eliminate assessment bias that offends or unfairly penalizes test takers because of personal characteristics such as race, gender, or socioeconomic status

5 Construction and improvement of selected response and constructed-response test items

6 Scoring of students' responses to constructed-response tests items, especially the distinctive contribution made by well-formed rubrics

7 Development and scoring of performance assessments, portfolio assessments, exhibitions, peer assessments, and self-assessments

8 Designing and implementing formative assessment procedures consonant with both research evidence and experience-based insights regarding such procedures' likely success

9 How to collect and interpret evidence of students' attitudes, interests, and values

10 Interpreting students' performances on large-scale, standardized achievement and aptitude assessments

11 Assessing English Language Learners and students with disabilities

12 How to appropriately (and not inappropriately) prepare students for high-stakes tests

13 How to determine the appropriateness of an accountability test for use in evaluating the quality of instruction

The current study adopts the assessment literacy definition presented by Popham (2011), which states assessment literacy is an individual's understandings of the fundamental assessment concepts and procedures deemed likely to influence educational decisions both in the classroom (i.e., classroom assessment) and those that impact the inside and outside of the classroom (i.e., accountability assessment). Popham's definition of the construct involves understanding of the following: (1) Assessment concepts, and (2) Contextual procedures that impact decisions. The first part is analogous to existing definitions outlined by Xu and Brown (2016) emphasizing knowledge of assessment terminology and concepts. The second part of Popham's (2011) definition contains the appropriate application of that knowledge to influence educational outcomes. Specific to the second half of the definition, assessment literacy involves the proficiency of the teacher to apply concepts and procedures (i.e., from the first half of the definition) to influence educational decisions within a particular context. Additionally, assessment literacy also entails how teachers select, introduce, and interact with assessment inside their classrooms.

The above definition of lists and skills, presented by Popham (2011) and used by the MAC is just one example of the lists presented when outlining and defining assessment literacy. These existing assessment literacy definitions are composed of a list of topics, concepts, and skills created by taskforces or organizations to outline what an educator needs to know about assessment to be assessment literate. Another such list, which is of primary focus for this study, is the *Standards for Teacher Competence in Educational Assessment of Students* from the American Federation of Teachers (AFT), National Council on Measurement in Education (NCME), and the National Education Association (NEA):

- Standard One: Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.
- Standard Two: Teachers should be skilled in developing assessment methods appropriate for instructional decisions.
- Standard Three: The teacher should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.
- Standard Four: Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
- Standard Five: Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.
- Standard Six: Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.
- Standard Seven: Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information (AFT, NCME, & NEA, 1990; para 13-30).

The *Standards* were written in order to communicate the importance of classroom assessment. These seven *Standards* focus on both classroom-based competencies and the role of the teacher in decisions beyond the classroom, as well as the teacher's participation in decisions related to classroom, district, state, and national assessments. The *Standards* were developed within a framework that follows the progression of these roles held by the teacher, starting inside the classroom and working outward toward the wider educational community.

The framework outlines the expectations for assessment knowledge and skills that a teacher should possess in order to perform well in five areas of activities including: (1) Prior to instruction, (2) During instruction, (3) After the appropriate instructional segment, (4) Associated with a teacher's involvement in school building and school district decision-making, and (5) Associated with a teacher's involvement in a wider community of educators (AFT, NCME, & NEA, 1990). Additional examples of frameworks used to develop assessment literacy lists and definitions exist and will be outlined in the paragraphs below.

Existing Assessment Literacy Frameworks

As was illustrated above, several definitions and manifestations of assessment literacy have been generated from assessment literacy frameworks, which outline the structure of the concept. Generally, assessment literacy frameworks have centered on classroom assessment practices largely used by teachers. This is true in the case of the *Standards* written by AFT, NCME, and NEA (1990). In particular, these frameworks address gaps in the assessment literacy of pre-service and in-service teachers. One such framework focusing on pre-service teacher assessment literacy was proposed by Siegel and Wissehr (2011). Their framework focuses on classroom principles of assessment for learning, and teacher knowledge of assessment tools and purposes.

Similarly, Gareis and Grant (2015) constructed a teacher-focused framework categorizing assessment literacy into three aptitudes for teachers and administrators. Their three suggested domains include: (1) Types of measures, (2) Quality of measures, and (3) Results and their uses (Gareis & Grant, 2015). Recently, Kahl, Hofman, and Bryant (2012) suggested the Assessment Literacy Domain Framework which builds on existing standards written by various educational institutions. The authors recommend assessment literacy mechanisms that emphasize employing standards to promote the use of results informing practice, programs, and measurement design. Each of these frameworks places the teacher in the center of the assessment process using a variety of skills including assessment, measurement, and results.

Other assessment literacy frameworks also include professional development paradigms, which is a departure from the mechanical list of skills that a teacher should possess to be assessment literate. Professional development paradigms in assessment literacy were developed because many teachers are not trained to be assessment literate, and therefore in-service and new teachers need to be prepared to understand assessment (DeLuca & Klinger, 2010; Popham, 2009; Volante & Fazio, 2007; Wang, Wang, & Huang, 2008). Professional development related to understanding assessment is also crucial as assessment is constantly changing. Inbar-Lourie's (2008) social-constructivist perspective of assessment literacy stated the need for assessment literacy through educator professional development. This was specific to teachers in language teaching contexts, but remains true for teachers regardless of content area as exhibited in various frameworks. Relative to these teachers, the components of assessment literacy suggested by Inbar-Lourie’s (2008) framework indicate that the definition of assessment literacy varies according to content area or role. The author articulated that assessment literacy for language educators would serve as the foundation for assessment training relative to these specific teachers. This example is important because it indicates that assessment is tied to context, and therefore each teacher may require a differing “set” of assessment knowledge.

Similarly, other teacher-focused frameworks propose the need to create an assessment literacy professional development pathway encompassing all the stages of teacher education and development (Xu & Brown, 2016). The authors proposed an assessment literacy framework (i.e., Teacher Assessment Literacy in Practice [TALiP]) consisting of five components: (1) Teacher conceptions of assessment, (2) Institutional and socio-cultural contexts, (3) TALiP, the core concept of the framework, (4) Teacher learning, and (5) Teacher identity (re)construction as assessors. These five stages of assessment, at the teacher level, align with the three domains of assessment knowledge that a teacher must have to be assessment literate. TALiP outlines these three domains of assessment literacy to include: (1) Educational assessment knowledge (i.e., the practical classroom assessment knowledge), (2) Knowledge of the interconnectedness of assessment, teaching, and learning (i.e., the intersection of personal perspective and theory), and (3) the assessor’s identity. Two of these three domains are similar to many of the other frameworks presented in this literature review. However, it is worth emphasizing that this framework acknowledges the context of the teacher (i.e., their identity), which aligns with the role of subject and content area as mentioned in Inbar-Lourie's (2008) framework.

The frameworks detailed above focus on the aspects of classroom assessment most traditionally conceptualized when educators hear the term "assessment." However, other components such as measurement and data must also be considered. Measurement concepts are amply displayed throughout the lists of skills suggested in the previous assessment literacy frameworks. The recognition of measurement as a distinct concept in classroom assessment was first highlighted in Popham's early work when testing, measurement, and assessment were somewhat interchangeable (Daniel & King, 1998; Popham, 1995).

Daniel and King (1998) illustrated an example of such empirical work that conflates measurement and assessment. They asked teachers about their familiarity with basic measurement principles, using key terms such as reliability, content validity, predictive validity, correlation, range, criterion related, mean, median, mode, and standard error. Furthermore, the questionnaire required teachers to make applied judgements about these concepts (i.e., interpreting correlations coefficients). This is not an isolated example, as basic measurement principles have been grouped with assessment literacy competencies across a variety of articles (Boyles, 2006; Brookhart, 2001; Gareis & Grant, 2015; Lambert, 1991; Popham, 2009; Taylor, 2009). The presence of measurement, as suggested by these studies and frameworks, may be inseparable from assessment in the classroom and therefore a part of what teachers need to know.

While knowledge of these concepts is beneficial, most frameworks suggest that teachers may benefit from a “working” (i.e., non-theoretical) understanding of measurement principles. Brookhart (2001) addressed this issue by identifying that her research questions attempted to measure teacher assessment literacy, but were in fact based on the knowledge of measurement principles. Brookhart (2011) further developed a set of assessment literacy principles that capitalized on this notion. This study illustrates the need for practical measurement knowledge for teachers. As both Brookhart (2011) and Popham (2011) have emphasized, teachers only need specific measurement knowledge to possess assessment literacy. Given this body of research, measurement training for teachers can be implemented in teacher education programs as a means to prepare teachers to understand the basic measurement concepts they will encounter in the field, such as percentiles.

However, technical knowledge such as measurement principles have even been measured by assessment literacy tests or inventories such as in the Assessment Literacy Inventory (ALI; Mertler & Campbell, 2005) and the Assessment Knowledge Test or AKT (Wang, Wang, & Huang, 2008). These teacher-focused assessment literacy measures explicitly question a teacher’s knowledge of measurement principles such as those listed in the previous paragraph. These measures, including measurement knowledge, emphasize the need for teaching basic and relevant measurement concepts to teachers, as they contribute to a teacher’s overall assessment literacy (Mertler & Campbell, 2005; Wang, Wang, & Huang, 2008). A theoretical understanding of measurement concepts, like error, is not necessary for teachers (Brookhart, 2011).

Therefore, there is an intersection between the proposed assessment literacy knowledge for teachers (e.g., Brookhart, 2011; Popham, 2011) and the recent measures of assessment literacy – the latter of which includes measurement knowledge. This intersection indicates the balance and need to incorporate basic measurement skills in teacher assessment literacy.

Other frameworks have stressed the importance of data and professional learning related to data in the process of assessment and assessment literacy. This subset of assessment knowledge is referred to as data literacy, which again stems from the idea that a literate individual knows how and when to apply basic knowledge and skills appropriately within a specific area. Supovitz's (2010) framework for data-related professional learning includes four processes in order: (1) Data capture, (2) Meaning-making, (3) Information sharing, and (4) Knowledge codification. Teachers are key in this process with regards to the classroom. Teachers are the individuals who are capturing data and carrying out much of the process. Additionally, Jimmerson and Wayman (2015) expanded on this framework by adding that individual learning and organizational learning support each other. They suggested that effective data-related learning includes what the individual learns are well as the group. For teachers, this illustrates that the assessment knowledge he/she gains in the classroom plays a larger role in the educational community and has the potential to benefit and benefit from collaborating with other teachers in learning to manage data.

Associated with data-related learning in assessment literacy is understanding uses of data and the processes involved in data use (i.e., what occurs in interacting with assessments, test scores, and other data forms). Coburn and Turner (2011) proposed a framework for data use which is comprised of the following components: (1) Process of data use, (2) Organizational and political contexts involved in data use, (3) Interventions to promote data use, and (4) Potential outcomes of these uses. The processes of data use are centered within the organizational and political context. The context, in turn, is influenced by interventions to promote data use resulting in outcomes such as organizational change, alterations to practice, and/or modifications to student learning. The presence of data, again, places teachers within a much broader context of assessment.

Data literacy, much like the previously-discussed assessment literacy frameworks, has been mainly teacher focused. Gummer and Mandinach (2015) constructed a data literacy framework for teachers which includes three domains: (1) Disciplinary content knowledge and practices, (2) Pedagogical content knowledge and practices, and (3) Data use for teaching knowledge and skills. This framework overlaps with the same domains as many others – content knowledge, pedagogy, and classroom practices. However, now the role of data is considered.

Lastly, Athanases, Bennett, and Wahleithner (2013) proposed a framework of systematicity in teacher inquiry which informs data literacy. The authors stated that data literacy for teaching incorporates the framework elements of data collection events, analysis, and using the information for teaching. This mimics many of the teacher-focused assessment literacy frameworks presented thus far with the evidence being collected as data. Given this emphasis on data, the model presents five steps in the process of the framework: (1) A data collection event, (2) Analysis, (3) Information for use in teaching, (4) Synthesis, and (5) Teacher knowledge that develops from reflecting on rounds of collection and analysis. This list illustrates similarities amongst the other frameworks, but the words, information, evidence, and results are becoming synonymous with assessment literacy.

These frameworks, coupled with empirical research results, highlight the increased need for a unified definition of the construct of assessment literacy. However, none of the existing frameworks or definitions identifies the full range of competencies that exists within assessment literacy. This is partly because, as has been suggested, the scope of assessment literacy may depend on the teaching context. A math teacher may not need the same assessment literacy knowledge base and skills as a language teacher.

Assessment literacy extends beyond the classroom and is not specific to only classroom-based decisions. This is consistent with many frameworks and definitions such as Popham's (2011), the *Standards* written by AFT, NCME, and NEA (1990), and the MAC, to name a few. Teachers are the educators in the classroom, but their involvement with assessment extends all the way to legislation. For this reason, this study of assessment literacy encompasses the full array of assessment knowledge and skills categories (e.g., classroom assessment, accountability assessment, measurement/data literacy) that are becoming increasingly important and unavoidable for pre-service teachers, in-service teachers, and teacher preparation programs (e.g., Popham, 2011). Given this all-encompassing definition of assessment literacy skills and knowledge, the self-efficacy or confidence in ability of pre-service teachers must be discussed.

Self-Efficacy

In the early 1960s, attention to inner experience, to internal processes, and to self-beliefs, such as self-efficacy, became the effort of many educators and psychologists (Maslow, 1954). However, inconclusive and inconsistent results lead psychology professionals to question this approach, and thus education, pedagogy, and practice also followed suit. Several methodological constraints in measuring latent traits like confidence provide the majority of evidence for these inconsistent results. Today, a resurgence of self-process research and theory has been connected to high-stakes educational contexts such a state accountability. For example, districts can choose indicators of self-process, efficacy, and emotion in reporting improvements in student performance. In relation to teachers, a recent study conducted by the Northwest Evaluation Association (NWEA; 2016) reported that the majority of teachers feel they are prepared in assessment or are assessment literate; however, they have little confidence in specific assessment-related abilities. At the high end of teacher confidence in assessment abilities, 56% of teachers felt prepared to administer assessments. Other skills such as selecting, developing, interpreting, communicating results, and informing practice received lower teacher confidence ratings ranging from 33% to 44% (NWEA, 2016).

More research and support for teachers in developing assessment-related confidence is needed, as is a common definition of the construct of assessment self-efficacy or assessment confidence. According to Bandura's Social Cognitive Theory, self-efficacy beliefs are influenced by the choices one makes and the courses of action one pursues (Bandura, 1977). This means that when an individual feels competent and/or has confidence in completing a task, he/she will choose to engage in it. On the other hand, when an individual feels incompetent and lacks confidence in a task, he/she will avoid engaging in it. Applying Badura's (1977) definition to the context of this study, it is vital that teachers feel competent (i.e., assessment literate) and confident (i.e., assessment confidence) in their assessment abilities in order to engage in the process.

Not only do competence and confidence influence engagement with a task, but self-efficacy also determines the amount of effort exerted on a specific task (Schunk, 1981; Schunk & Hanson, 1985; Schunk, Hanson, & Cox, 1987). The amount of effort includes a measurable amount of time spent on a task, and also an individual's ability to face adversity or overcome obstacles experienced during a task. Overall, a greater sense of self-efficacy yields more chances of persisting to complete a task given obstacles (James, 1985). Another consideration within the construct of self-efficacy is its influence on outcomes. When an individual believes that he/she will successfully achieve an outcome, he/she will have more confidence in their abilities (Bandura, 1986). Therefore, self-efficacy should be approached as a multi-faceted construct, requiring not only confidence and competence, but attention to effort and application.

The role of experience and task-specific performance must also be considered in teachers' assessment-related self-efficacy. Individuals develop self-efficacy from previous experiences, and these occasions help shape self-efficacy when an individual encounters a similar task (Pajares, 1996). Schunk (1996) defines these confidence judgements, which are self-reported scales measuring confidence levels, as self-efficacy for performance because self-efficacy depends on and influences actions involved in a specific task. However, when faced with a related, novel task, self-efficacy for performance is replaced with confidence judgments labeled self-efficacy for learning (Shunk, 1996). Using these related concepts, teachers that develop early self-efficacy in training (i.e., self-efficacy for learning) may also demonstrate an increased self-efficacy for performance in practice, as new assessment-related tasks emerge. Furthermore, the combination of self-efficacy for learning and self-efficacy for performance is important for institutions of higher education that need to consider how previous experience and performance will shape teacher confidence as he/she evolves throughout their careers and experiences with assessment.

While no such construct of "assessment confidence" has been defined and researched, theorists such as Bandura (1997) have accounted for self-efficacy's contextualization. According to Bandura (1997), self-efficacy is context-dependent, indicating that an individual's self-efficacy beliefs are grounded in a task type or the performance domain in question. Additionally, the context in which self-efficacy is measured can vary according to task, difficulty, and various situational circumstances that can influence self-efficacy (Bandura, 1997). Another social cognitive theorist (Marsh, 1993) stated the contextualization of self-efficacy is an essential consideration given that multidimensional self-efficacy ignores domain-specific knowledge. Both Marsh (1993) and Bandura (1997) assert that due to this contextualization, measures of self-efficacy should be context-specific (i.e., not global appraisals), as aggregate measures ignore domain-specific information.

Additionally, research has detailed the relationship between positive and negative self-efficacy involving any task. The idea of positive and negative self-efficacy refers to the varying levels of self-efficacy any individual can have in relation to a specific task. As Pajares (1996) outlined, an individual may have lower self-efficacy on a given task because that task is not meaningful to them. For example, he/she may take no pride in accomplishing that task. Additionally, lower self-efficacy often exists when an individual knows he/she is “bad” at performing a certain task, or he/she knows the task is outside of his/her ability level. The inverse could also be true in that an individual may feel high positive self-efficacy for a task he/she performs poorly on because of his/her perspective on the given task. The individual may also have higher self-efficacy if a task is too easy for them, and they are aware of the low level of difficulty. A null level of self-efficacy is also possible, where an individual truly does not feel positive or negative self-efficacy because he/she is unaffected by the result of the task.

Based on Bandura’s model, Dembo and Gibson (1984) created the Teacher Efficacy Scale containing several contextual factors related to teacher confidence. However, these factors were not related to pedagogical contexts like area of teaching practice. The nature of de-contextualized teacher self-efficacy and recommendations for increasing self-efficacy among teachers have been outlined (Dembo & Gibson, 1985). Other existing studies on teacher self-efficacy have focused on generalized feelings of self-reported confidence. Walters and Daughtery (2007) defined the construct of teacher self-efficacy as any individuals’ judgments or beliefs regarding his or her ability to accomplish critical instructional tasks. In this survey study, the authors reported relationships between teacher experience and self-efficacy and mastery and self-efficacy. Thus, teacher self-efficacy is noted to be particularly important for new teachers and those that teach complex content areas.

Lastly, a link between self-efficacy and teacher belief in their responsibilities in the classroom was reported by Brookhart, Loadman, and Miller (1994). This represents another conceptualization of teacher self-efficacy; however, it does not take into account skills like assessment. The authors framed “assessment” as responsibilities where a teacher may or may not feel strong confidence in their skills. Again, the link between confidence and competence was demonstrated as teachers who perceived something as a high-priority responsibility were more likely to feel increased self-efficacy when they had a solid understanding of the task.

The majority of teacher self-efficacy studies focus on specific content domains and pedagogical areas like math, science, technology, and foreign language teaching (e.g., Bursal & Paznokas, 2006; Demo & Gibson, 1985; Garbett, 2003). This reflects the assertion that a teacher must possess high self-efficacy, specifically in the knowledge of their specific domain of teaching. Such studies have shown that teacher confidence is relative to specific content areas within a teaching specialty. For example, Watson (2001) studied math teacher confidence across nine areas of math knowledge including assessment-related aspects like basic descriptive statistics. The results indicated that teacher confidence in math education and pedagogy significantly varied across grade levels. In other words, high school math teachers reported higher confidence in relation to higher-order content knowledge such as basic statistics. The disparity between confidence and measured competence (i.e., content knowledge) was greater across grade levels (i.e., Early Childhood, Middle Childhood, and Adolescent Education) than across knowledge of content and skills (Watson, 2001). This indicates that there was a larger difference between perceived competence and confidence when exploring the differences between elementary school teachers and high school teachers. For example, when investigating the difference between teachers' knowledge of a specific skill, like addition and subtraction, and confidence in said skill, elementary school teachers who consistently teach the properties of addition and subtraction were more confident, but not necessary more competent.

Similar studies have shown the connection between self-efficacy and the teaching of science and technology (e.g., Bursal & Paznokas, 2006; Garbett, 2003; Graham, Burgoyne, Cantrell, Smith, St. Clair, & Harris, 2009). Garbett (2003) administered a self-report confidence and competence measure to a group of Early Childhood Education teachers. The results revealed a significant relationship between how confident teachers felt in a specific science content domain and how competent they were in their knowledge of the domain, such as biology, chemistry, and physical science. However, complementary results of a science multiple-choice test indicated that teachers overall were not as competent as they believed themselves to be.

Graham and colleagues (2009) also conducted a study of science teacher confidence by using the Technological Pedagogical Content Knowledge (TPCK) survey. The authors found that over the course of an eight-week program, as competence in specific areas of science and technology teaching increased, confidence also increased.

The results indicated that classroom teachers developed content-specific confidence as they learned. Finally, Bursal and Paznokas (2006) found a positive correlation between teacher self-efficacy in science and math. This suggests that the confidence required to teach these pedagogical content domains might be related. Across all studies, more confident teachers generally have a greater understanding and belief in their content knowledge.

The majority of the teacher-self efficacy research outlined above focuses on in-service teachers, which highlights the need for research on self-efficacy in pre-service teacher populations. Studies like Graham and colleagues (2009) indicated that confidence increases as teachers learn specific content knowledge and skills. One study on pre-service teacher confidence reported an increase in confidence before and after their practicum experience (Main & Hammond, 2008). This is consistent with previous research that suggests experience plays a role in confidence (Pajares, 1996). In addition, the specific domain assessed by Main and Hammond (2008) was classroom management. Although not a content domain, classroom management is comprised of a unique set of skills and knowledge that teachers are expected to learn in their teacher education programs.

To summarize, teachers in practice who are confident are those who are also competent in their respective content domain. Confidence is context-dependent and related to the specific pedagogical content domain of the teacher. The link between confidence and content knowledge was illustrated in studies such as Watson (2001) that provided evidence of confidence and content knowledge differences in various grade levels of math teachers. Evidence from two studies suggests that there is a positive relationship between experience and confidence. Interestingly, these studies report that the length of experience and its impact on confidence was shown in as little as eight weeks (Graham et al., 2009) in student teaching practicums (Main & Hammond, 2008). Considering the findings summarized here, the relationship between pre-service teachers' confidence and content knowledge are seldom represented, emphasizing the need to conduct research on content knowledge and confidence in specific skills and domains in this population.

Summary

From the objectives in Chapter 1, this study aims to examine the relationship between pre-service teacher assessment literacy and confidence in assessment. More specifically, the two objectives are: (1) Examining the psychometric properties (i.e., reliability and validity) of the newly-modified CALI including a confidence measure for pre-service teachers, and (2) Investigating the impact of pre-service teacher assessment literacy and assessment confidence on performance assessment scores (edTPA). Considering these two objectives, this chapter reviewed the existing literature related to both assessment literacy and confidence/self-efficacy.

The process of assessment was presented and summarized to illustrate the practice and capacity in which teachers use assessment (Wiliam & Black, 1996). Then, this chapter reviewed the various definitions of assessment literacy and the suggested content knowledge and competencies teachers need to be assessment literate. Topics such as data literacy as a component of assessment literacy were addressed but are not a main component of assessment literacy in this study as literature and previous empirical research need continued development. While teachers do not need a statistician's level of measurement understanding, there are basic measurement principles related to assessment that are suggested for teachers to be assessment literate (e.g., Brookhart, 2011; Popham, 2011).

Finally, the literature review presented Bandura's (1977) Social Cognitive Theory, which provides the foundation for understanding confidence in this context. Several studies emphasized the principles of Bandura's (1977) theory in presenting confidence in teaching (e.g., Hoy & Spero, 2005; Tschannen-Moran & Hoy, 2001). Research on assessment literacy has focused on in-service teachers based on various existing assessment literacy standards. The above literature review provides the initial considerations for investigating confidence and competence in pre-service teachers. The following chapter (Chapter 3) details the methodology of this study.

CHAPTER III

METHODOLOGY

Purpose

Research on the assessment literacy of pre-service teachers has grown since the instatement of the No Child Left Behind Act (NCLB, 2002; Brookhart, 2011; DeLuca & Klinger, 2010; Gareis & Grant, 2015; Inbar-Lourie, 2008). With this legislation, the frequency and duration of standardized assessment has also increased across the country. Currently, with the Every Student Succeeds Act (ESSA), educators are once again faced with the same issue – the centrality of assessment and accountability in measuring student academic improvement and teacher and institutional effectiveness (DeLuca & Bellara, 2013). Thus, the role of the teacher continues to extend far beyond that of pedagogical skills and/or content knowledge.

Teachers are tasked with being conduits of assessment at the classroom, local, state, and national levels. Not only are teachers responsible for administering the assessments, they also are charged with explaining results to students, their parents, and various stakeholders (Kahl, Hofman, & Bryant, 2013). These results not only impact classroom grades, but influence district and state level accountability systems (DeLuca & Bellara, 2013). As many definitions of assessment literacy have noted, the inclusion of accountability assessment in teacher's responsibilities extends far beyond traditional classroom assessment knowledge (e.g., Popham, 2011).

The rapid evolution of assessment has changed how it is conducted and reported. For instance, there has been an influx of student growth percentile reporting in lieu of other gain score or ranking techniques (Betebenner, 2009; Blank, 2010; Walsh & Isenberg, 2015). For most people, when understanding concepts like student growth, there can be a steep learning curve involving the connection between statistics and measurement and assessment-related changes. This component of assessment literacy is subsumed under Popham's (2011) definition of accountability assessment. Newer concepts that describe student improvement, such as growth percentiles, require some working knowledge of statistics and measurement. These statistics go far beyond traditional reports of proficiency and percentiles (Betebenner). Teachers must understand, to a certain degree, what this aspect of assessment means for them and for their students. The connection between a concept such as student growth and teachers can be seen in Popham's (2011) definition of assessment literacy used in this study. Assessment literacy also contains an individual's understanding of how this aspect of assessment influences the teacher outside of the classroom. The increasing complexity of assessment and its changes may, in turn, impact teachers' confidence in their abilities to understand, conduct, and report standardized and classroom assessment.

Additionally, other aspects of assessment literacy impact the classroom such as test item writing and creation, administration, reporting, and teacher evaluation systems (Brookhart, 2001). Considering the many facets of assessment literacy related to teachers and previous research on pre-service teacher assessment literacy, researchers must investigate and target how to directly influence and increase teachers' assessment literacy

(Popham, 2009; Taylor, 2009; Wang et al., 2008). As in-service teachers across various grade levels and content areas, identified by Mertler (2003), report struggles with assessment-related understanding and practices, higher education and teacher training are considered the “starting point” to remedy teachers’ assessment literacy knowledge deficits.

This study had two main research objectives: (1) to investigating the psychometric properties of an assessment literacy and confidence measure for pre-service teachers, and (2) to examine the relationship between an assessment literacy and confidence measure and pre-service teacher performance outcomes. More specifically, the first objective was to investigate the psychometric properties (i.e., reliability and validity) of the Classroom Assessment Literacy Inventory (CALI; Mertler, 2003), using a sample of undergraduate university students in Early Childhood, Middle Childhood, and Adolescent teacher education programs at a large Midwestern university in the United States. The first research question stated: “What are the psychometric properties of the newly developed assessment literacy and confidence measure for pre-service teachers?” Following an item-level examination of CALI, a subsequent question was also posed: “What is the internal structure of the modified CALI?”

Specifically, this study examined the aspects of assessment knowledge and understanding (i.e., assessment literacy) within a sample of pre-service teacher undergraduate students by investigating the psychometric properties (i.e., content and construct validity, internal consistency reliability) of participant scores on the CALI. Assessment knowledge was measured using a traditional diagnostic/summative approach (i.e., the CALI) as well as the performance-based model used by the edTPA. Evaluating participants using both methods produced information on classroom-based knowledge and applied knowledge or ability. In this case, pre-service teacher undergraduate students were those in their final semester of coursework where they typically participate in student teaching and licensure examinations. Research has been conducted on pre-service and in-service K-12 teachers since the CALI's creation using this specific measure (or similar versions), in which some reliability evidence has been demonstrated (Mertler & Campbell, 2005). However, further psychometric examination was warranted, as over ten years have elapsed since the CALI's creation.

Additionally, this study included a newly-developed measure of confidence after each question on the CALI (i.e., termed "assessment confidence"). Along with an examination of pre-service teachers' assessment literacy levels, this study examined how confident these teachers were in their assessment knowledge. To reiterate, this version of the CALI (i.e., the "modified CALI") refers to the original CALI multiple-choice questions written by Mertler (2003) along with the confidence questions after each CALI assessment-related content-matter question. Thus, the psychometric properties (i.e., reliability and validity) of the modified CALI were the primary focus of the first objective, and a secondary emphasis was on the construct of assessment confidence and how it relates to assessment literacy (i.e., assessment content knowledge) and dimensions within the CALI.

The second objective of the current study was to investigate the impact of assessment literacy and assessment confidence on performance assessment outcomes.

The second research question states: “What is the impact of assessment confidence on the relationship between pre-service teachers' assessment literacy and performance assessment outcomes?” All graduating students within the proposed sample were required to take the edTPA performance assessment as a graduation requirement. At this time, the edTPA was not a requirement for licensure in the state of Ohio. However, due to the increasing national use of the edTPA exam, previously known as the Teacher Performance Assessment, the proposed study evaluated the relationship between classroom assessment knowledge, assessment confidence, and high-stakes performance assessment outcomes. Exploring the connection between assessment knowledge and assessment confidence and edTPA performance could prove beneficial for programs and a state moving towards edTPA licensure requirements as it provides information related to pre-service teachers’ understanding of assessment concepts as well as confidence in these concepts. These students may have a solid understanding of assessment, but are not yet confident in their understanding, which can lead to implementation of additional opportunities to apply assessment skills to bolster their confidence.

The current study bridges the gap between teacher understanding of assessment and confidence in assessment knowledge within the context of teacher preparation and education programs. Using Rasch Analyses including Principal Components Analysis (PCA), Confirmatory Factor Analysis (CFA), and Moderated Multiple Regression, this study investigated the relationship between the components within assessment knowledge and how confident teachers are in these aspects of assessment. This not only identifies the strengths and weaknesses related to the assessment literacy of pre-service teachers embarking on their first few years in the classroom, but it also provides insight into the assessment preparation of students within teacher education programs.

Context

The university and sample used for this study were from the state of Ohio and are held to this state's teacher licensure requirements. For this reason, a discussion of the state's definition of a licensed teacher in the state of Ohio is outlined. An Ohio teaching license has high reciprocity at the national level. However, while a teaching license from Ohio can be used across a number of states, the state does not grant reciprocity to teachers with existing licenses from Louisiana, Montana, Missouri, Nevada, New Mexico, North Dakota, Texas, and Wyoming (Ohio Department of Education [ODE], 2017). The state of Ohio requires graduates of in-state accredited teacher education programs to complete three steps in order to obtain the initial 4-year Resident Educator License (ODE). These steps include: (1) Completing a Bachelor's degree, (2) Obtaining proper certificates, and (3) Taking required assessments. While all states require a Bachelor's degree, in Ohio, each state-approved teacher preparation program has its own curriculum and coursework. Most curricula in the state incorporate subject mastery and basic pedagogical theory and practice. Given the combination of curriculum and practical fieldwork, teacher education programs typically include all requirements for the Ohio Teacher Certification Program. This means that upon graduation, students have completed all required examinations, coursework, and classroom experiences to be licensed teachers as defined by the state of Ohio.

In general, curricula across teacher education programs generally emphasize foundational knowledge and skills, pedagogy, and educational technology (Ohio Higher Education Educator Licensure Program Standards and Requirements). These components are what create commonality across teacher education programs; however, no two programs are the same. Some programs may decide to emphasize different curricular aspects. Moreover, most all curricula nationwide aim to prepare students to plan, instruct, and implement student learning related to their field of pedagogical study. Applying these practices often occurs in a fieldwork requirement. For instance, the state of Ohio requires fieldwork experience for licensure, which includes student teaching. Fieldwork can also include observations and internship roles in addition to student teaching (ODE, 2017). The combination of this experience allows students to apply what they have learned in their teacher preparation program.

Ultimately, all teacher education programs in the state of Ohio must be approved by the ODHE (www.education.ohio.gov). The approval of the program, a student's completion of that program, and the integration and completion of classroom fieldwork experience, like student teaching, comprise two-thirds of the requirement in the state. Finally, in order to gain licensure in the state, a candidate must complete any assessments required by the state. These assessments typically assess pedagogy and knowledge as well as subject-specific knowledge. At this time, the state of Ohio requires teachers to pass the Ohio Assessment for Educators (OAE) pedagogical assessment. Ohio also requires the appropriate OAE Content Assessment or Praxis Subject Assessment for the area of licensure. The use of these two exams provides evidence for the teacher's basic skills and knowledge and content area abilities according to the state of Ohio. However, it should be noted that the edTPA also achieves both of these objectives as it assesses pedagogical ability according to content knowledge domains. However, the edTPA is a performance assessment whereas the existing exams, the OAE and OAE Content Assessment, are traditional assessments. Ohio is taking steps towards implementation of the edTPA as is the trend across the nation.

University X's teacher education program includes: Early Childhood Education (Grades K-3), Middle Childhood Education (Grades 4-9), Adolescent and Young Adult Education (Grades 7-12), and teacher education programs for Health Education, Art Education, Music Education, Special Education, Physical Education, Teaching English as a Second Language, and World Languages (University Website). However, for this study, these specialized programs were also included and of interest since each of these teacher preparation programs require attention to assessment and the edTPA. Like most major state universities, University X's teacher education programs require students to take a general Educational Psychology course, which addresses basic assessment fundamentals and concepts. Additionally, in their second year of course work, students are enrolled in content-specific pedagogy courses that being to address concepts of assessment at the classroom level in relation to the following specialties: Early Childhood Education, Math Education, Integrated Language Arts, Integrated Sciences, Integrated Social Studies, Biology, and Chemistry, etc.

Each of the groupings of teacher education programs (e.g., Early Childhood, Middle Childhood, and Adolescent Education) requires classroom experience prior to graduation. This is also true for other programs such as Art, Music, and Foreign Language teaching; however, each program implements fieldwork differently. Classroom experience is fulfilled through a combination of classroom observations or internships and student teaching. On average, teacher education students in these programs have the opportunity to spend at least two semesters in the classroom via observation, internship, and student teaching (University Website). These classroom experience requirements are satisfied throughout several semesters of student coursework, culminating in student teaching during the final semester. Additionally, prior to graduation, students must submit their edTPA performance-based assessment, which is used throughout the US to emphasize, measure, and support the skills and knowledge that all teachers need to be independent in the classroom. The edTPA assessment is not yet required for licensure in the state of Ohio, although steps toward implantation have begun according to the state policy page on edTPA.org. The assessment, however, is a graduation requirement at University X, which currently requires completion but does not have a designated passing score and a licensure requirement in dozens of states nationwide.

Participants

Results from this study generalize to teacher education program students who are approaching graduating. This population represents pre-service teachers who are about to begin their careers and the level of assessment knowledge possessed immediately upon exiting a typical teacher education program. These pre-service teachers have yet to fully apply their knowledge of assessment and assessment practices in the classroom outside of supervised teaching (i.e., student teaching). This population has minimal experience reporting, analyzing, or discussing high-stakes assessment results in relation to their own classrooms or schools. Additionally, these graduating students have limited practice administering and/or communicating assessment results to parents either at the classroom level and/or the district, state, and national levels.

There are two samples in this study. The sample for the pilot phase and the sample for the second, confirmatory phase are outlined below. As this study focuses on psychometrics and measurement development, the discussion of both samples is presented.

Pilot Sample. The sample in the initial pilot phase of this study was comprised of 165 second- and third-year teacher education students in Early Childhood, Middle Childhood, and Adolescent Education across the content areas of English/Language Arts, Integrated Sciences (i.e., biology, chemistry, physical science), Mathematics, and Integrated Social Studies, as well additional teaching specialties (i.e., Health, Physical Education, Art, Music, Special Education, and Foreign Language). These participants were in either their second or third year of study, as defined by university credit hour requirements, and were recruited from intact classes during the Spring semester of 2017. This sample received the CALI, with all original items, and the additional confidence scale.

Second Phase Sample. The sample in the second phase of this study was comprised of 112 fourth-year teacher education students from the same programs as the pilot sample. Fourth-year student were defined as those who were about to graduate during the semester these data were collected. Specifically, these students were student teaching and taking the edTPA during this final semester in their degree programs. This sample received the CALI after ten items were eliminated based on pilot data analyses. This sample also received the confidence scale questions and edTPA scores from these participants were collected.

Similarities and differences exist between these two samples (i.e., the pilot sample and second phase sample), and the purposeful decision to use distinct second-/third-year and fourth-year student groups for data collection is discussed in the following paragraphs. First, these two samples were chosen due to their similarities in coursework progression and completion. These two groups would have completed, or been in the process of completing, the same coursework and therefore had the most similar level of assessment knowledge. At minimum, all students would have taken the general Educational Psychology course, which focuses on assessment, and likely entered their content-focused methods course(s), as well as been exposed to basic assessment principles. Additionally, most students had some form of classroom experience, ranging from passive observations or internships to student teaching.

The main difference to consider between the second-/third-year and the fourth-year groups is student teaching. While most second-/third-year students had some level of classroom experience (i.e., observing, guest teaching, assisting), during the second phase data collection in Spring 2017, all fourth-year students had completed their student teaching requirements in preparation for graduation at the end of the semester. This suggests that fourth-year teachers would have likely had the opportunity to apply some of their assessment knowledge in the class. It also suggests the presence of this applied knowledge coupled with what was taught via the curriculum coursework.

Measures

The CALI is a 35-item measure of classroom assessment literacy developed by Mertler (2003; Appendix A). The version of the CALI used throughout this study is the first version developed by Mertler (2003) including 35 multiple-choice questions, which have been made free and available online. The 35 questions are applied in nature and require participants to respond to either factual knowledge items or read short classroom explanations or scenarios before selecting a response. As was mentioned in Chapter 2, these questions align with The *Standards for Teacher Competence in the Educational Assessment of Students* (AFT, NCME, & NEA, 1990). Since there are seven standards and 35 questions, there are five questions grouped under each standard (Appendix B). That means the questions in the CALI are associated with the benchmarks and the skills suggested for assessment literate teachers as defined by these *Standards*.

The Classroom Assessment Literacy Inventory (CALI)

The CALI is unique from other existing assessment literacy measures in its use of multiple-choice questions and its classroom-based approach. Two measures employing similar approaches and alignment with the same *Standards* exist, the Assessment Literacy Inventory (ALI; Mertler & Campbell, 2005) and the Teacher Assessment Literacy Questionnaire (TALQ; Plake, 1993; Plake, Impara, & Fager, 1993). The structure of the CALI and its use of multiple-choice questions was selected for this study because of its continued development, evidence of reliability and validity that will be outlined below, and practicality of administration and time. For instance, the ALI (2004) is comprised of five vignettes each followed by seven multiple choice questions. If a student fails to understand or read one vignette, it can impact all seven questions, whereas in the CALI, each question requires a new application of knowledge.

Next, the definition of the construct "assessment literacy," as defined by the CALI and its developers, is discussed. According to Mertler and Campbell (2005), assessment literacy is an educator's ability to recognize sound assessment, evaluation, and communication practice by understanding the purposes associated with different assessment methods and student achievement, communicating assessment results effectively, and using assessment to maximize student motivation. While assessment literacy may appear unified in this definition, the construct encompasses several domains of knowledge such as classroom assessment, measurement, evaluation, statistics, and psychometrics. The CALI aligns these domains within the *Standard*s list discussed below. Often skills and knowledge associated within these varying domains yields slightly different definitions of assessment literacy as other definition have been proposed by Brookhart (2001), Plake (1993), Popham (2009), and Stiggins (1995).

The definition assumed by Mertler and Campbell (2005), mentioned above, was operationalized by a set of standards or expected competencies for teachers, which further reflects the scope of skills encompassed in assessment literacy (i.e., *Standards for Teacher Competence in the Educational Assessment of Students*; AFT, NCME, & NEA, 1990). This set of standards was written to define a common threshold for teacher competence in student assessment and to guide teacher training and preparation. There are seven *Standards* which cover a wide scope of assessment skills encompassed by the domains previously listed. The seven standards are:

1. Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.
2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions.
3. The teacher should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.
4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.
5. Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.
6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.
7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information (AFT, NCME, & NEA, 1990; para 13-30).

However, some opposition to these *Standards* has since been acknowledged. The *Standards for Teacher Competence in the Educational Assessment of Students* (AFT, NCME, & NEA, 1990) are arguably outdated. Several key authors in the field have suggested similar, more current standards of assessment literacy knowledge for teachers

(Brookhart, 2001; Popham, 2009; Stiggins, 1999). Nonetheless, Mertler and Campbell's (2005) argument for the relevancy of the *Standards* (AFT, NCME, & NEA, 1990) remains valid due to the general overlap and commonalities between existing standards lists. The *Standards* are still widely cited and considered relevant across the field of assessment and measurement in education (Reynolds, Livingston, Willson, & Willson, 2010).

Authors compared a more recent set of standards written by Stiggins (1999) to the *Standards* (AFT, NCME, & NEA, 1990) and emphasized the extensive overlap comparing both lists. The lists contain the same general skills and focus, but organize and emphasize different aspects according to context. For example, the *Standards* are entirely focused on teachers. The structure of domains and skills within each list differs slightly, but the overlap between both sets of standards is present. For example, Stiggins (1999) includes the competency of applying proper assessment methods which is also listed in the *Standards* as competencies one and two (AFT, NCME, & NEA, 1990). This overlap suggests that the fundamental skills and knowledge associated with the construct has not drastically changed over time.

Many domains and skills are encompassed in assessment literacy warranting an examination of the unidimensionality of the CALI measure. The field of assessment research and teacher preparation has yet to establish if different constructs, for example measurement knowledge, are distinctly separate dimensions within assessment. However, Mertler and Campbell (2005) identified competing constructs when they acknowledged the reason for creating both the CALI and the ALI. The CALI includes shorter applied and context-related multiple-choice questions compared to the ALI; however, both measures have a strong emphasis on application. While the overall reliability of the scores on the measures is comparable, Mertler and Campbell (2005) believed that some level of classroom assessment knowledge may only exist in the classroom. The authors suggested that pre-service teachers lack this experience and knowledge because they have yet to be in front of their own classroom. This level of hands-on classroom knowledge, referring to experience in the classroom, could therefore impact assessment literacy as measured by the CALI; however, evidence of this has yet to be demonstrated.

Given this description of the construct and scope of the measure, the primary uses of the CALI are twofold: (1) It provides a mechanism for educators to measure assessment literacy, and (2) It informs decision-making and guides practice (Mertler, 2003). This is especially important for programs approaching licensure changes such as those created by edTPA's implementation and focus on assessment. Up to now, the CALI has been utilized primarily as a measure of pre-service teachers' assessment literacy. Early versions of assessment literacy measures, such as the TALQ, were used with in-service and pre-service teachers and produced scores with lower or no reliability evidence. Through the CALI's focus on pre-service teachers, there is also a potential third use or extension of the intended second use of the measure – program evaluation. The CALI provides a benchmark for how an assessment curriculum at an institution is preparing its teacher education students (Mertler, 2005). By assessing student performance in relation to the *Standards*, a model for adjustments in teacher education programs and curricula can be made.

The CALI's scores have shown evidence of reliability and validity according to its first official use as a measure in Mertler's (2003) study. The evidence of reliability of the measure, prior to this study, came from the reliability estimates of its predecessor, the TALQ. Mertler (2003) adapted the TALQ based on its evidence of reliability with in-service teachers (r_{kr20} = .54; Plake, Impara, & Fager, 1993), and with pre-service teachers (r_{kr20} = .74; Campbell, Murphy, & Holt, 2002). In Mertler's (2003) study, the data from CALI responses from an in-service group of teachers reported an internal consistency reliability estimate of .74 and .54 for pre-service teachers.

The above reliability estimates did not follow traditional procedures for evidencing reliability of scores (Crocker & Algina, 2006). Instead, Mertler (2003) used the test-retest paradigm but on separate samples. This was assumed, but not explicit, as the author collected data from two samples, in-service (*N* = 197) teachers from all grades and content levels and pre-service (*N* = 67) secondary education teachers. Furthermore, the internal consistency reliability reported by Mertler (2003) was for the overall measure, which can be problematic if the measure has underlying dimensions (Cortina, 1993). The current study explored the reliability of the scores and unidimensionality of the CALI, and also investigated the presence of underlying dimensions and their individual reliability estimates using Rasch Analysis and Confirmatory Factor Analysis (CFA). Mertler and Campbell (2005) conducted Rasch Analysis with the ALI, but not with the CALI.

The existing evidence of the CALI's score reliability was supported by a literature review outlining reported psychometric properties (i.e., reliability and validity). Mertler

(2003) reported a KR-20 reliability value of .74 to .75, which according to Nunnally (1978) was acceptable reliability for a measure of this size, with this purpose, and level of associated risk. However, this threshold of reliability only occurred within the in-service teacher sample. The group of pre-service teacher responses on the CALI had lower reliability (r_{kr20} = .54). Given other existing research and evidence of reliability in this sample using the ALI and TALQ, the magnitude of the reliability index was not a deterrent in using this measure. All of this information must be taken into consideration along with the purpose of the measure. If the measure was being implemented for high-stakes purposes, such as teacher salary and accountability, the reliability coefficient should be closer to .90 (Nunnally, 1978). Most high-stakes tests are intended to distinguish those who have mastered the material from those who have not. Therefore, the purpose of the measure and the intended use of the results must be considered. For the present study, the stakes might not be as high; however, the use of this measure and the possible implications it has on edTPA preparation present an argument for future high-stakes use.

Through the various developments of the CALI, Mertler (2003) and Mertler and Campbell (2005) indicated that their expertise validated the alignment between the questions and the standards. Since the *Standards* incorporate a set of behaviors and tangible skills, the authors used these as the criterion for measuring the validity of this measure. Secondly, the authors acknowledged the process of item review with attention to best practices for item writing such as clarity, wording, and language suggested by Clark and Watson (1995). It was unclear if outside experts were consulted during this process; however, the process was iterative and depended on consensus regarding item appropriateness and quality. Lastly, there was evidence of the process of criterion validation outlined by Cronbach and Meehl (1955) throughout the development of the CALI as well as the ALI.

While more information about the psychometric analysis of this measure is needed, an issue of importance is validity. The measure relies strongly on a set of standards which represent a construct with several factors. Therefore, as with all measurement development in general, continual construct validity evidence is needed (Clark & Watson, 1995). To validate the alignment between the measure and the *Standards*, a CFA can be conducted to evidence the internal structure. In order to proceed with this process, studies should provide evidence of the psychometrics when applying any changes to a measure. Collecting additional data to conduct CFA should also be considered, as CFA fit indices are susceptible to sample size issues. However, samples which are too large have a tendency to reject any possible model (Marsh, Balla, & McDonald, 1988). Lastly, concurrent validity between this measure and participant performance in their assessment coursework could be examined extending the use(s) and purpose(s) of the measure.

Overall, based on general measurement standards as presented by Crocker and Algina (2008), the psychometric properties (i.e., reliability and validity) of the scores on the CALI are provided with some evidence of reliability and validity to assess pre-service teacher assessment literacy. This information and research was used in selecting this measure for the current study. While the ALI and TALQ are both variations of this same questionnaire, the CALI had distinct advantages for data collection with the sample in this study. Additionally, with other researchers improving the CALI and evidencing the measure's performance, Mertler (2003) continues to add to a growing area of research seeking to validate the scores on such measures of assessment literacy in order to improve teacher training.

The Assessment Confidence Scale

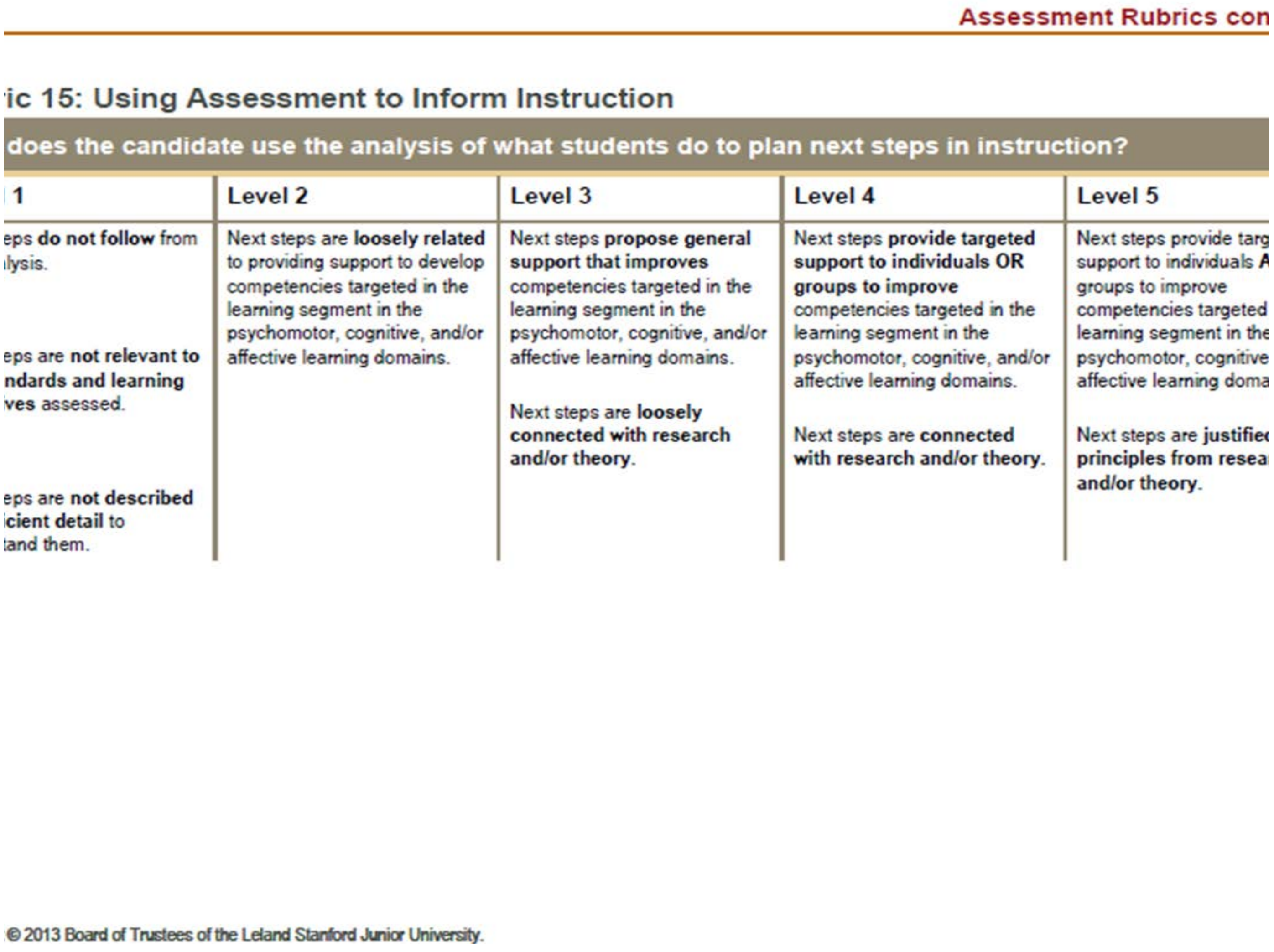
The confidence scale used in this study was created with the intention of measuring polytomous data, which are data that are measured on a scale with greater than two categories (i.e., dichotomous). The newly-created assessment confidence scale consisted of one question repeated 35 times throughout the CALI after each item. The question was: "How confident are you in your answer to the above question?" All questions appeared in a subsequent vertical format. The participant read and responded to one CALI content question, and then he/she scrolled down to the next question immediately vertical to the CALI question, and responded with his/her level of confidence. Each confidence question consisted of five response options. The response options were presented horizontally from the least amount of confidence to the most amount of confidence with a neutral option in the center. The response options were as follows: "Completely Unconfident," "Mostly Unconfident," "Neither Confident Nor Unconfident," "Mostly Confident," and "Completely Confident." More information on the confidence scale will be outlined in the following sections detailing the development of the modified CALI.

The edTPA

The edTPA is a performance-based assessment of educator preparation developed by the Stanford Center for Assessment, Learning, and Equity (SCALE) and recently acquired by Pearson Education, Inc. Field tests in 2013 and continued evaluation using data from 2014 and 2015 provided validation of the scores produced on the exam and evidence of its score reliability as reported by SCALE. The edTPA is the first large-scale performance assessment of its kind. It is not only designed with the novice teacher in mind, but also the objective of providing data to teacher preparation programs that support curricular evaluation and change. Conceptually, it aligns with a variety of widely-accepted state-level teacher licensure standards such as college and career readiness benchmarks, the Interstate Teacher Assessment and Support Consortium (InTASC) standards, and major teacher evaluation frameworks. According to edTPA.com, the exam is currently used by over 700 institutions and in more than 30 states. In some cases, such as Ohio, the state does not yet require the test for licensure. However, some universities in Ohio are using edTPA as a graduation requirement for pre-service teachers because the state is taking steps toward total state-wide implementation for all new teachers via the state policy page reported by etTPA.org. Therefore, the edTPA may potentially serve as the licensure exam in the majority of, over 30, states across the country.

As a performance assessment, the edTPA requires pre-service teachers to engage in a variety of tasks aligned with what is expected as an in-service teacher. This means that the students do not select answers on a traditional paper-and-pencil standardized exam, but rather demonstrate knowledge of certain skills and concepts. The edTPA is both a measure of basic pedagogical skills and subject-specific knowledge. This aligns with the licensure requirements of many states that necessitate a new teacher demonstrate basic skills and content-specific knowledge. The edTPA is used in these capacities, by both states and higher education institutions, to measure and support the skills and knowledge that all teachers need in their first day in the classroom. There are more than 27 content-specific exam options ranging from Pre-Kindergarten to 12th grade teaching specialties as of 2017 (edTPA.com). In some states, students are required to take their specific exam upon graduation and in order to obtain licensure. For instance, in this study's sample, a participant intending on graduating in the Spring of 2017 with a Bachelor's Degree from the Adolescent Education Program with a focus on Math would take the edTPA Math test for high school teachers prior to their May graduation date.

Although the exam is offered in 27 different content domains, all versions of the edTPA assess the same three tasks: (1) Planning, (2) Instruction, and (3) Assessment. The planning section contains the following five rubrics: planning for content understands (e.g., producing lesson plans), knowledge of students, supporting academic language development, and planning assessments. Instruction contains the video recorder portion of the exam assessing: learning environment, engaging students, deepening student learning, subject-specific pedagogy, and analyzing teaching effectiveness. Lastly, the assessment rubrics measure analyzing student learning, feedback (i.e., involving two rubrics), analyzing students' academic language understanding and use, and use of assessment to inform instruction. An example of an assessment rubric (i.e., Assessment Rubric 15, using assessment to inform instruction) is show below in Figure 1. Within these three teaching-related performance domains, two additional skills relevant to practice were embedded in the exam. These additional skills include academic language and analysis of teaching.



Assessment Rubrics con

ic 15: Using Assessment to Inform Instruction

does the candidate use the analysis of what students do to plan next steps in instruction?

1	Level 2	Level 3	Level 4	Level 5
eps **do not follow** from lysis. eps are **not relevant to ndards and learning ves** assessed. eps are **not described icient detail** to tand them.	Next steps are **loosely related** to providing support to develop competencies targeted in the learning segment in the psychomotor, cognitive, and/or affective learning domains.	Next steps **propose general support that improves** competencies targeted in the learning segment in the psychomotor, cognitive, and/or affective learning domains. Next steps are **loosely connected with research and/or theory**.	Next steps **provide targeted support to individuals OR groups to improve** competencies targeted in the learning segment in the psychomotor, cognitive, and/or affective learning domains. Next steps are **connected with research and/or theory**.	Next steps provide targ support to individuals **A** groups to improve competencies targeted learning segment in the psychomotor, cognitive affective learning doma Next steps are **justifie principles from resea and/or theory**.

© 2013 Board of Trustees of the Leland Stanford Junior University.

Figure 1. The edTPA Assessment Rubric 15.

Measuring these components requires a three- to five-day documented learning experience. This consists of three to five lessons constructed and planned by the candidate. The candidate must submit evidence that documents all stages of the process of lesson construction and implementation. Examples include authentic lesson plans, materials, rationales, and reflections to provide insight into his/her process of teaching and learning as educator. Ultimately, students also video record themselves teaching in the classroom to provide evidence of their instructional capabilities. All of this information is used to assess students across the three dimensions and two sub-dimensions noted above (i.e., Planning, Instruction, Assessment, Academic Language, and Analysis of Teaching).

The majority of content areas apply the types of evidence outlined above across fifteen total rubrics. Each of the three major domains measured have five rubrics, respectively (i.e., five rubrics for planning, five rubrics for instruction, and five rubrics for assessment). These five rubrics consist of five levels of performance ranging from one to five, where one represents not ready and five represents an accomplished novice teacher. In sum, a candidate receiving a score of five across all planning rubrics, would receive a raw score of 25 for their planning abilities. Therefore, planning accounts for one third, or 25 points, or the possible total score.

The progression of performance is detailed for each rater according to each content-specific exam. Raters are given rubrics that detail the progression, are calibrated, and instructed as to what each point in the rubric means. Each task is scored by two different raters. At this point, inter-rater reliability information for the edTPA exam was last reported as ranging from .83 to .92 (Stanford Center for Assessment, Learning, and Equity, 2013). This information is provided by distinguishing what determines an automatic rating of “one,” what type of response and response qualities qualify as a “two,” “three,” and so on. Furthermore, distinction between levels is included, such as information that helps a rater decide the difference between a four a five. Current raters are active or retired Higher Education Faculty within a state-endorsed teacher preparation program or a K-12 teacher or administrator with a valid license, and raters must have classroom experience within the last five years.

There are 25 points possible for each of the three domains noted above, and a total of 75 points for the entire edTPA exam. These scores are presented to the students in this sample electronically on a score report with various score components including total raw scores per section and rubric. The overall report is considered the summary score (i.e., including all scores reported). The report also provides students with rubric scores for each of the fifteen rubrics, rubric averages that summarize the scores across all five for each of the three domains, and a total edTPA score.

What constitutes a passing score is at the discretion of the state or institution. In many cases, this score also fluctuates according to the content specialty of the exam. For instance, a score of 35 is considered passing in California for teachers of Classical Languages like Latin, but a score of 41 is required for any single-subject elementary school teacher (edTPA.com). In the case of most universities in Ohio where these is no state-established passing rule, a score of 37 is considered passing according to the edTPA. The university sampled in this study does not have an official passing score at this time. Completion of the exam is considered passing. The edTPA's importance in this study is to indicate its role in accessing assessment knowledge of new teachers. Additionally, it provides insight into the curricular alignment of current teacher preparation programs adapting to the use and implementation of this exam, with a strong focus on applied knowledge.

Procedures

The initial development of the CALI is summarized in the following paragraphs, as the current study did not create the measure. Mertler (2003) used the TALQ (Plake, 1993; Plake, Impara & Fager, 1993) to create the CALI. The transition from the TALQ to the CALI included general sematic reconstruction. The purpose of the Mertler's (2003) development of the CALI was to improve the overall clarity of the TALQ by adjusting the applications in the questions to common names and words. The CALI contained the same content, 35-items, and continued alignment with the same standards in the TALQ.

Measure Development of the CALI

For the current study, the CALI was modified in three ways as described below. First, more demographic questions were added to the beginning of the modified CALI. In the initial version of the CALI (Mertler, 2003), there were only seven demographic questions. The modified CALI consisted of twenty demographic items. These items were used to gather more detailed information about participant age, gender, race, teacher education program, grade point average, student status (e.g., junior, senior, etc.), and socioeconomic indicators. Other information about student experience and exposure to assessment was collected. These questions focused on the number of courses or workshops in which students were exposed to assessment either as the sole topic or part of a comprehensive course. The original CALI also included two questions about the participants' perceptions of their teacher and assessment preparation, which were retained in the modified CALI.

Secondly, the CALI was re-organized to ask the demographic and perception questions first rather than last. In the initial CALI, all of the non-content related questions were posed at the end of the measure. If left to the end of the measure, many participants may not answer those items, which would provide no information about the demographics of participants who did not respond to the remaining items. Lastly, the confidence scale was added to the CALI. The confidence scale was a single question measuring the respondents' confidence in their answer to the previous content-related question. Thus, there were 35 content-based items and the confidence scale appeared one time after each of these items. Detailed psychometric analysis and further definition of these indices and results will be presented in Chapter 4.

Measure Development of the Confidence Scale

The creation of the confidence scale followed Clark and Watson's (1995) guidelines for developing scales measuring hypothesized latent traits. First, the latent trait was identified as confidence, and the purpose was to create a scale to assess how confident participants were in their responses to each question about assessment. The goal of this scale is to assess to what degree of confidence the participant felt in his/her response to the assessment question.

In this study, a Likert scale was used, as the construct was conceptualized to exist on an ordinal scale. This ordinal scale ranged from the most negative point, or the lowest level of confidence, and increased from left to right. That is, the right or final option ordered on the scale was the most positive response. The five response options were presented horizontally from left to right and read as: “Completely Unconfident,” “Mostly Unconfident,” “Neither Confident or Unconfident,” “Mostly Confident,” and “Completely Confident.” These points were coded on a scale of 0 to 4 to reflect the least amount of confidence as 0, and the greatest amount of confidence as 4, with a neutral option coded as 2. More specifically, participants read the question, “How confident are you in your response?” followed by the five ordinal response options. The 5-point Likert scale followed each content question on the CALI. This scale is consistent with existing confidence and self-efficacy research both in its construction and methodology including the sematic construction (e.g., “Mostly Confident”) and use of a 5-point Likert scale (e.g., Betz & Borgen, 2010; Maurer & Pierce, 1998; Sander & Sanders, 2003).

Data Collection

All data were collected in the Spring of 2017 across a sample of students in various teacher preparation programs. Prior to data collection, all permissions and approvals were granted via the university’s Institutional Review Board (IRB; see Appendix C). For the pilot phase of this study, the data were completely anonymous. No identifiable information was collected, and the recruitment and collection were organized through key-personnel involved with this study.

The same procedure for IRB approval and participant recruitment was conducted for Phase 2 of this study, which took place immediately after the pilot (Appendix C). The key difference between Phase 1 and Phase 2 was the use of a temporary identifier. In order to connect Phase 2 student responses to edTPA scores, FERPA-approved email addresses, which are directory information, were used to temporarily record survey responses. An administrator at University X was the only individual to see the identifiable email addresses and student edTPA scores. Upon linking the edTPA scores and survey responses, the temporary email address identifier was permanently deleted. The students were offered no compensation for participating and informed they could exit the survey at any time. Non-verbal consent was issued to all participants and appropriate IRB approval waived the need for a physical signature at the time of consent. Waiver of consent via signature was used to maintain the anonymity of the responses and refrain from collecting any information that could be used to link responses back to the students at any point.

The CALI

Across both phases of this study, the CALI was administered to students using Qualtrics. The modified CALI was uploaded into the online Qualtrics program and numerically coded to record correct responses. Students in both phases received an email sent to their university email account. During Phase 1, all students received the same link as the CALI was anonymous and there was no need for any identifiable information. A copy of the email communication is included at the end of this document (see Appendix C). Students were given two weeks to begin and complete the survey and responses to all questions were required. After two weeks, participants received a reminder email to complete the survey. In some cases, the researcher, or a member of key personnel on this study, supervised large-scale collection of data. This occurred when the faculty member allowed a researcher to attend a class meeting and escort students to a computer lab to complete the CALI. These groups ranged from five to 25 participants. This increased the participation rate and promoted the faculty members and program engagement in this study.

For Phase 2, the modified CALI link was sent to each individual email address using Qualtrics. This means that each participant had a unique link according to their email address. All CALI participation and communication during Phase 2 was via email. The individual links and the use of password-protected email accounts met appropriate identification validation procedures, and provided assurance that students were completing the CALI that was unique to their email address. By using unique links, it was possible to track the email address and the responses that corresponded to it. In the Qualtrics output, the responses were recorded along with the link and email address to which the link was sent. Once the necessary information was obtained, all identifiable information was permanently removed. As in Phase 1, participants received an initial email and then a two-week reminder (see Appendix C). A final reminder was sent nearly four weeks after initial contact. Due to the nature of this data collection and the Qualtrics software, reminder emails were only generated and sent to those participants who had not completed the survey at that point.

Once students received the email with the CALI link, they were instructed to click the link to the CALI when they had time to complete the survey, which on average lasted 20 minutes. When students clicked the link, they were taken to an initial screen that presented a summary of the informed consent, a pdf attachment of the entire informed consent, and the instructions to complete the modified CALI. All participants were aware that by clicking the start arrow, they were consenting to participate and could exit the study at any time.

The CALI consisted of two groups of questions, as was outlined previously in the measure development section. Participants completed one group of demographic questions before beginning the modified CALI questions. Participants then completed the second group of questions containing the CALI items. There were 35 questions for Phase 1 participants and 25 questions for phrase 2 participants. Lastly, participants saw a "Thank You" and exit screen with the contact information of the researchers should they have questions. At this point, responses were recorded in Qualtrics and participants could close the window or browser.

The Assessment Confidence Scale

All confidence data were collected along with the CALI responses as noted the in previous paragraphs. That is, the confidence scale was included with the CALI questions, which were the second group of items in the online CALI administration. The confidence scale was part of the modification to the CALI. This modification, as was outline above, included a 5-point Likert scale response for each question. This response scale appeared immediately after each CALI question. For example, when participants completed CALI question one, coupled with the question was the confidence scale asking, “How confident are you in your response?” Participants selected their level of confidence from the 5-point Likert scale and proceeded to CALI question two.

The edTPA

The collection of edTPA scores was conducted through University X’s College of Education. Because these scores are not freely and openly available, permission was obtained via the IRB application for phrase 2 of this study (see Appendix C). The Phase 2 modified CALI responses were collected and downloaded from the Qualtrics program. From this information, an Excel spreadsheet was created including the student’s email address (i.e., the temporary identifier) and responses. This document was shared with key personnel on this study who also serves as an administrator at University X. This individual then used the email address to identify participants. Columns were inserted into the Excel document and the corresponding edTPA information was added. This included overall performance and performance across the three sub-scales of the edTPA (i.e., Planning, Instruction, and Assessment). Following this, email addresses, which were the only identifying information, were removed before the linked data were viewable by the lead researcher. Only the Director of Assessment and Accreditation, whose position included access to these data, saw the students email address and score concurrently.

Data Analysis

The following section outlines the data analysis techniques that were used on the Phase 1 data in this study.

Phase 1 Pilot Sample

Each major phase of this study required a different analytical technique due to the goal of the analysis related to the research objective.

Rasch analysis of the modified CALI. In accordance with the first research objective of this study, Rasch Analysis was used to examine the psychometric properties (i.e., reliability and validity) of the modified CALI using the Winsteps® 4.0.0 computer program (Linacre, 2017). Rasch Analysis, as part of the Item Response Theory (IRT) family, was developed to overcome some of the problems and assumptions associated with Classical Test Theory (CTT). IRT does not require sampling assumptions, nor does it require normal distributions, which is ideal for different item structures such as those on the modified CALI (i.e., dichotomous and polytomous). Using CTT with non-normal data would require using Medians in lieu of Means, Interquartile Ranges (IQR) instead of Standard Deviations, and Spearman Correlations instead of Pearson Correlations. Additionally, IRT does not require that measurement error be considered the same for all individuals' responses on a measure (Bond & Fox, 2015; Wright & Stone, 1979).

Specifically, the Rasch Analysis (based upon IRT principals) was expected to provide information on the range of difficulties presented by the items, the category usage based on the rating scale, and any potential underlying factor structure. The Rasch model was used in place of other IRT models because it is the most parsimonious model for the development of measures given the present sample size (Wright, 1997). The Rasch model fulfills the requirements of fundamental measurement, that of sample-free instrument calibration and instrument-free person measurement, where more sophisticated IRT models fall short (Andrich, 2004). Incorporating Rasch Analysis is expected to provide information on the psychometric properties (i.e., reliability and validity) of the newly-developed instrument and insight into the item behavior.

Many IRT models are available, and the simplest one is the Rasch (i.e., one parameter, or 1PL) model. Rasch Analysis allows for the creation of an interval scale for both item difficulty and person ability. Rasch considers both how the items perform relative to other items in the sample, as well as how persons perform in relation to the sample and the difficulty of the items. Rasch scores are reported in logits which are placed on the same scale that measures person ability and item difficulty (Andrich, 2004; Wright & Stone, 1979). The Rasch model calculates the probability that a person will get an item correct and that an item will be answered correctly by a person. If the probabilities are different from the observed data, the results indicate that the data do not fit the model (i.e., using fit statistics; Wright & Stone, 1979).

Chief among Rasch Analysis indices includes item reliability, separation, fit, and category structure or thresholds. Each of these concepts provides insight into the instrument’s psychometric properties (i.e., reliability and validity) as part of the measure development process. In Rasch Analysis, item reliability represents the ability to replicate item placement based on the Rasch estimates (Bond & Fox, 2015). Values closer to 1 represent strong reliability, while values closer to 0 indicate almost no certainty in replicating item difficulty estimates. Separation refers to the variation in item difficulties within the instrument, where larger values suggest good distribution of difficulties (de Ayala, 2009). Separation values that are less than one typically represent overlap or redundancy of items at a certain difficulty level. Reliability and separation are critical to consider, as are item fit statistics.

Rasch Analysis also includes an examination of item and person fit statistics (i.e., infit and outfit) to address problematic responses. Examination of the infit and outfit indices assist in identifying poor fit between the data and the Rasch measurement model. Infit and outfit statistics represent discrepancies between responses, with infit being weighted by values close to the expected value of difficulty or ability, and outfit being unweighted, leading it to be more sensitive to outlying responses (de Ayala, 2009). Infit violations can appear when easy items are not endorsed by capable persons, whereas outfit violations typically occur when the level of difficulty is outside of the pattern of responses such as a lucky guess on a difficult question (Linacre, 2000). Infit values are reported as Mean Square values (MNSQ) and as standardized *z*-values (ZSTD). MNSQ statistics show the amount of distortion present with 1.0 as the expected values. Values less than 1.0 indicate observations that are too predictable, with values greater than 1.0 indicating unpredictability (Linacre). ZSTD values show the unlikelihood that the data fit the model with 0.0 as their expected values. Indices less than 0.0 indicate an item that is too predictable with 0.0 indicating unpredictability, or the lack of information needed for an item to determine an individual’s ability. Standardized values can be either positive or negative. It is essential that each item demonstrates adequate fit with the construct, as evidenced by fit statistics that represent correspondence with the Rasch model.

Additionally, the analysis considers the response option usage and category structure thresholds. This is specific to polytomous data and will be used for the confidence scale, and not the dichotomous content question, since the confidence scale uses five Likert points. The thresholds (i.e., step calibrations) are the difficulties estimated for choosing one response category over another (e.g., how difficult it is to endorse “Completely Unconfident” or instead of “Completely Confident”). The step calibrations should increase monotonically (i.e., have ascending threshold values), and the distance between threshold values should be neither too close together nor too far apart on the logit scale. Bond and Fox (2015) suggest thresholds should increase by at least 1.4 logits to show distinction between categories, but no more than 5 logits (i.e., to avoid large gaps in the variable). The thresholds will be analyzed to provide further evidence related to the category structure within the model and the use of response options on the instrument.

In the current measure, there were two response types on the modified CALI. The content questions focusing on assessment knowledge were dichotomous. Responses to dichotomous items were simply correct or incorrect. A person with a high score on the CALI is said to have an increased level of assessment knowledge. On the other hand, the responses to the confidence scale questions were ordered categories (i.e., a Likert scale) from “Very Unconfident” to “Very Confident.” This indicates increasing levels of a response on the variable of interest (i.e., assessment confidence). Much like the dichotomous items, a person with a higher total score is said to show more of the variable assessed. Given the two different response types, these data were analyzed separately – one set of analyses for the dichotomous assessment knowledge items and another set of analyses for the confidence scale data. The Rasch Dichotomous Model was used for all binary data responses and the Rasch Rating Scale Model (RSM) was implemented for the Likert-scale confidence responses (Andrich, 1978).

For the current study, approximately 50 to 100 participants were needed for preliminary analysis of the modified CALI if item calibrations were to be stable within ± 1 logits (i.e., 99% CI - 50 people), or ± 1/2 logits (i.e., 95% CI - 100 people; Linacre, 1994). The present study collected responses from 165 pre-service teachers at the initial pilot stage. At this time, the amount of data and the suggested sample size for a Rasch Analysis of a measure at these beginning stages had been met. The Rasch Analysis of the modified CALI produced several indices of item fit as outlined above. These fit statistics, mainly infit and outfit, were examined to determine items or persons that were problematic for model fit. The MNSQ fit statistics were inspected if they exceed 1.2, which is suggested of high-stakes measures (Wright et al., 1994). In this study, the CALI was analyzed as a high-stakes measure because of its potential use and connection to edTPA outcomes. The higher the MNSQ fit statistic, the more questionable the information (Wright & Stone, 1979). Based on this information, items were eliminated to produce the best possible Rasch model, and the remaining items were comprised of the newly-revised modified CALI that can be used in future analyses and measure development/refinement studies. No outlying persons were identified following the same criteria outlined in this section.

Rasch Principal Components Analysis (PCA). Continuing to investigate the first research objective, the psychometric properties (i.e., reliability and validity) of the modified CALI, required Rasch PCA. Rasch PCA, just like all Rasch analytical techniques, investigates the differences between what the Rasch model predicts and what is observed. PCA uses residuals to explore patterns in the data. Items with similar patterns typically share a substantive attribute and therefore create a dimension or component. PCA of these residuals identifies characteristics shared in common among items. This analysis uses an inductive approach using standardized residuals to uncover the structure of the measure (Linacre, 1998). PCA often provides indications of secondary structures or sub-dimensions within the measure. PCA is not appropriate to use for testing hypotheses or theories, but only to explore and describe relationships among groups of items. Thus, no formal hypotheses will be made other than that the PCA yields an interpretable component, or groupings of items, based on the responses to the items within this sample. Since PCA is somewhat exploratory in nature, no inferential statistical processes were used; however, the second analysis of the data in Phase 2 included a confirmatory analysis of any component structure rendered.

When conducting a PCA, the goal is to reduce the dimensions present in the data. PCA is used to reduce the amount of variance present in a large set of variables into a smaller set of variables that accounts for the majority of the information (Linacre, 1998). Mathematically, PCA transforms the numbers of possible correlations between variables into a smaller number of uncorrelated variables called principal components (Wright, 1996). The first principal component accounts for the largest possible amount of variability in the data, with each subsequent competent continuing to account for as much variance as possible. Each extracted component represents a different construct in the measure. In PCA, each item in the analyses receives a factor loading value, which are the correlation coefficients between the variables (rows) and factors (columns) in the correlation matrix (Wright). These factor or component loadings are the key to understanding the underlying nature of particular dimensions as items are grouped into components according to the loading values (Bond & Fox, 2015). Additionally, PCA eigenvalues are used to express the amount of variance accounted for by the data. The eigenvalue of any present component should be close to 2 in order to determine the presence of a strong dimension within the items (Bond & Fox).

In this analysis, PCA was utilized to serve the purpose of reducing items into components with as little a loss of information as possible (Thompson, 2004). It was anticipated that some of the items on the modified CALI would be identified as problematic, but the results would serve to inform the reduction of items into fewer, more manageable components. Because PCA is based on the assumption of perfect reliability of the data (Fabrigar, Wegener, MacCallum, & Strahan, 1999), the analysis was employed to assume and account for all of the variance among the items (Thompson, 2004). Additionally, using PCA as a method of analysis allows for the inclusion of all possible variability among the responses. Thus, it is helpful in determining the most useful items to be investigated in further analyses and development of the measure.

PCA was used in this instance instead of traditional Exploratory Factor Analysis (EFA) for several reasons. PCA is different from traditional EFA, though PCA is often confused with and misused as a substitute or variant of EFA (Henson & Roberts, 2006). Research suggests that in some cases, PCA serves as a first step of item screening, prior to subjecting items to EFA (Matsunaga, 2010). Therefore, PCA was seen as a necessary first step to reduce the items into fewer, more manageable components using all available variance within the items. Conducting a PCA is often used to summarize the information available provided by a given set of variables (i.e., items) and reduce it into a fewer number of components (Fabrigar et al., 1999). The intent of employing PCA was to use the subsample to determine components through item reduction, which would inform the confirmatory analysis used later. Ultimately, the results of the PCA were expected to provide a framework of components that could be validated using CFA.

Phase 2: Confirmatory Study

Factor Analysis is a statistical procedure used to find a small set of unobserved variables (i.e., constructs, latent variables, or factors) that can account for the covariance among a larger set of observed variables (i.e., indicators). There are two types that underlie the broad statistical family of Factor Analysis. Exploratory Factor Analysis (EFA) is a data-driven approach that aims to discover the factor structure of an instrument. Confirmatory Factor Analysis (CFA) is a theory-driven approach that aims to confirm hypothesized factor structures (Dimitrov, 2013). For the purposes of the proposed study, CFA will be conducted in this phase in order to provide confirmatory evidence of the factors explored in the PCA analysis (i.e., in the first phase).

CALI Confirmatory Factor Analysis (CFA). Factor Analysis techniques were used to investigate the second research objective. The following proposed analysis investigated the second phase of data collected for this study which consisted of 112 responses from graduating teacher education students. The first phase of this study analyzed data using Rasch PCA. Given these components, the second phase of analysis included "traditional" Factor Analysis, which is a statistical procedure used to find a small set of unobserved variables (i.e., constructs, latent variables, or factors) that account for the covariance among a larger set of observed variables (i.e., indicators). Based upon the structure established within the pilot testing phase, the results of the full-scale administration were analyzed using factor analytic techniques. The results of the analyses were used to evaluate fit with the preliminary factor structure and to assess the underlying factor structure of the instrument. The intention of the factor analytic approach is to validate the PCA results. An examination of the factor structure, loading values, and inter-item correlations was conducted to determine the dimensions of assessment literacy captured by the items in the instrument and are presented in detail in the next chapter.

One type of Factor Analysis is Confirmatory Factor Analysis (CFA). CFA is a theory-driven approach that aims to confirm hypothesized factor structures (Dimitrov, 2013). As a hypothesis-testing analysis, CFA is used when the goal of the analysis is to affirm the validity of a hypothesized model of factors and their relationships to a set of observed variables (Briggs & Cheek, 1986; Dimitrov, 2013). This happens by specifying various restrictions on the factor model based, in this case, on the results of the Rasch PCA, and testing the residual matrix to determine whether it still contains significant covariation (Schumacker & Lomax, 2012).

Assumptions. In order to conduct CFA, three multivariate assumptions must be met. These assumptions include: (1) Normality, (2) Linearity, and (3) Homoscedasticity. Normality is observed when the distribution of each observed variable (i.e., the CALI items) is normal (i.e., univariate normality). Secondly, Linearity is met when the joint distributions for all combinations of observed variables are normal (i.e., multivariate normality). Lastly, Homoscedasticity occurs when all the bivariate scatter plots are linear and all points show equal variance (Dimitrov, 2013). However, these assumptions are only tenable for analyzing variables that are continuous. This study did not use continuous scores for any CFA analysis, but rather item level dichotomous scores (i.e., 0 and 1) for each item on the 25-item modified CALI. Therefore, any distributional assumptions about the observed variables cannot be made.

The presence of ordinal data (i.e., assessment confidence responses) must be addressed. In order to convert the data into an accepted format for CFA there are several possible correlational transformations: Kendall's *tau−b* correlations, Spearman's rho, and polychoric correlations. The polychoric correlation was first introduced by Pearson (1900) who expressed the relationship uses a contingency table. Pearson's (1900) contingency table was based on the assumption that there was an underlying normally distributed relationship between continuous variables. Therefore, ordinal data is representative of continuous data. Babakus, Ferguson, and Jöreskog (1987) demonstrated that polychoric correlations present the best results on the basis of squared error and factor loading bias. Additionally, Jöreskog and Sörbom (1996) indicated that the polychoric correlation, among six other possible correlations, is superior for evaluating the ordinal data when the underlying bivariate normality holds.

In order to calculate the polychoric correlation, it is assumed that the ordinal data are inherently connected to underlying continuous data. In this case, relationships between ordinal data can be measured with the help of related underlying continuous data. Based on this relation, polychoric correlations demonstrate relationships between ordinal variables. Múthen (1983) indicated the process of connecting these two types of data. Observed ordinal variables are related to unobserved continuously distributed variables (i.e., latent variables). Assume that x is the observed ordinal variable with m

categories and $x*$ is the underlying continuous variable. According to Múthen (1983) a monotonic relation is presented as:

$$x = c \Longleftrightarrow \tau_{c-1} < x* < \tau_c,\ c = 1, 2, ..., m, \qquad [1]$$

where

$$\tau_0 = -\infty,\ \tau_1 < \tau_2 < ... < \tau_{m-1},\ \tau_m = +\infty, \qquad [2]$$

τ. are thresholds categorizing continuous data into ordinal data.

Given this relationship, the assumption of the relationship between the ordinal variable and the underlying continuous variable must be checked. The assumption can be investigated in LISREL by examining the test of underlying bivariate normality. The null hypothesis of this test is that underlying bivariate normality holds and should be used for all ordinal data in this study. The results of this test will indicate if p values of all pairs of variables are larger than the standard criterion of .95 (Hooper, Coughlan, & Mullen, 2008). If this assumption is met, the calculation of polychoric correlation can be conducted using Olsson's (1979) procedure which expands from Múthen's (1983) original representation of the structure of ordinal data and its relationships.

Estimation methods. The purpose of model estimation is to minimize the differences between the sample covariance matrix and the model implied covariance matrix. Four popular estimation methods are: Maximum Likelihood (ML), Robust Maximum Likelihood (RML), Unweighted Least Square (ULS) and Diagonally Weighted Least Square (DWLS). Jöreskog (1969) presented Maximum Likelihood (ML) as an estimation method to illustrate a CFA model based on the assumption of Normality. ML is a method of full information estimation, which like Factor Analyses, allows for statistical inference like significance testing and goodness-of-fit evaluation. ML is typically used with normally distributed continuous data. For this reason, ML is often the default setting in most software programs. However, while ML can support slight departures from Normality, in practice the assumption of Normality is often violated (e.g., Hu, Bentler, & Kano, 1992). If the assumption is violated, then model results may not be reliable. ML can also produce chi-square and error bias when is it applied to skewed or non-normal data, which can impact the goodness-of-fit indices.

Considering these issues, ML is used as long as the data are normally distributed. If the data are not normally distributed, RML (Satorra & Bentler, 1994) and WLS (Browne, 1984) are two other estimation options. RML has the same estimation properties as ML, but it has modified standard errors and chi-square. RML is an estimation procedure based on the use of an asymptotic covariance matrix. The inclusion of this covariance matrix produces less biased standard errors and performs well facing different sample sizes and degrees of non-normality. WLS, on the other hand, is not recommended for the present research study as it requires large sample sizes to calculate its weight matrix (Jöreskog & Sörbom, 1996). In sum, ML is used if the data are normally distributed. If the data are not normally distribution, RML is considered as the best alternative. However, both ML and RML are best suited for continuous variables which are not used in the CFA analysis in this study.

The two estimation methods for ordinal data are ULS and DWLS (Brown, 2012; Yang-Wallentin et al., 2010). WLS can also be used for ordinal data, but still requires large sample sizes, and was therefore not used in this study. ULS and DWLS are similar to WLS but differ by using a weight matrix under the fit function. DWLS uses a weight matrix, which only contains the diagonal elements of the asymptotic covariance matrix, while ULS uses the identity matrix as its weight matrix. Previous research has shown that ULS outperforms DWLS and gives more precise estimation by means of less bias and smaller standard errors than DWLS (Forero et al. 2009; Rigdon & Ferguson, 1991). ULS is also recommended when a polychoric correlation matrix is used because it considers the weight matrix and non-convergence (Babakus et al., 1987). For this reason, ULS was considered for all CFA analyses of ordinal data representing the total item scores (i.e., dichotomous) to the modified CALI.

CFA sequence of steps. CFA follows a sequence of five steps: (1) Model Specification, (2) Model Identification, (3) Model Estimation, (4) Model Testing, and (5) Model Modification. These five steps will be outlined in the paragraphs below and applied to the assessment knowledge and confidence items in the modified CALI.

Model specification. Model specification is the construction of the theoretical model to create a covariance structure based on theory and prior research (Schumacker & Lomax, 2012). Model specification involves determining how many factors underlie the data, the factors that are related to the observed variables, which factors are expected to correlate and the errors that are expected to correlate, and which factor loadings should be held equal (Dimitrov, 2013). The goal of model specification is to determine the best possible model that generates a sample covariance matrix (*S*) that closely fits the population covariance structure (Σ). If the sample covariance matrix (*S*) of the specified model is not consistent with the population covariance matrix (Σ), which is generated from the population covariance structure, the model is considered misspecified (Schumacker & Lomax, 2012). In other words, in order to determine the best possible model the researcher must consider the underlying factors of their observed data, determine the best specifications for the relationships between these factors, and continue to specify these relationships until the data represented by their sample acceptably match the population model.

Model identification. The necessary condition in CFA is determining whether the hypothesized model is identified. A model is considered identified when every parameter is distinguished, and parameters represent any measurable characteristic of a population. Moreover, models with more information than unknown parameters are simply identified models and can be solved uniquely and tested statistically (Fox, 1983). The number of free parameters to be estimated must be less than or equal to the number of distinct values in the matrix *S* (i.e., sample variance-covariance matrix) in the model. Free parameters in the hypothesized model are factor loadings, measurement error variances, and correlations among the latent variables. The number of distinct (i.e., unique) values in the matrix *S* can be calculated using the formula:

$$p(p+1)]/2 \qquad [3]$$

where p is the number of observed variables in the model. If the number of distinct values in the sample matrix *S* is greater than or at least equal to the number of free parameters, the model is identified (Schumacker & Lomax, 2012).

Model estimation. In model estimation, the researcher finds the appropriate “fitting function” (Schumacker & Lomax, 2012) that helps to minimize the difference between the population covariance matrix Σ and the sample covariance matrix *S.* When elements in the matrix *S* minus the elements in the matrix Σ equal zero ($S - \Sigma = 0$), then the χ^2 will be equal to zero, which indicates a perfect model fit to the data (Schumacker & Lomax). Among the various fitting functions that are used in CFA, the Weighted-Least Squares (WLS) estimation method is the one recommended for dichotomous data, such as the assessment questions in the modified CALI (McDonald, 1970). The WLS estimation method generally requires a large sample size. It is considered an Asymptotically Distribution-Free (ADF) estimator, which does not depend on the normality assumption (Schumacker & Lomax, 2012).

LISREL (Jöreskog & Sörbom, 1993) statistical package for analysis of covariance structures was used to conduct CFA in the current study. With dichotomous variables, LISREL employs two steps in order to analyze a matrix of polychoric correlations rather than covariances. First, the thresholds and the polychoric correlation matrix are estimated using the maximum likelihood methods, and second, the inverse of the asymptotic covariance matrix (ACM) is estimated using WLS. The two steps can be applied using LISREL PRELIS command.

Model testing and modification. The aim of model testing is to determine how well the data fit the model. In other words, to what extent is the theoretical model supported by obtained sample data (Schumacker & Lomax, 2012). In CFA, if the fit of the model is good (i.e., Σ and *S* are similar), then the specified model is supported by the sample data. Otherwise, the specified model is not supported by the sample data, and model modification is needed to achieve better fit.

Modification is conducted by evaluating a series of indices that provide information on discrepancies between the proposed and estimated model. These modification indices provide information about covariances present in the data. In order to determine the fit of the specified model, the following fit indices will be reported: The Satorra-Bentler Scaled Chi-Square (i.e., for dichotomous data), the Root Mean Square Error of Approximation (RMSEA), the Standardized Root Mean Residual (SRMR), the Goodness-of-Fit Index (GFI), and the Adjusted Goodness of Fit Index (AGFI). The Satorra-Bentler Scaled Chi-Square test examines the similarity between the sample covariance matrix S and the produced model implied covariance matrix Σ. A non-significant result ($p > .05$) is desired. A good model fit is also evidenced by RMSEA ≤ .05, SRMR < .08, and GFI and AGFI > .95 (Dimitrov, 2013).

As a result of conducting CFA, the decision of whether the sample confirms the components found in the PCA analysis of the modified CALI will be made. Additionally, a better understanding of how assessment knowledge and assessment confidence function and whether it is suitable to be used in the pre-service teacher education context will be inferred. However, further analysis is recommended to support the use of the modified CALI on pre-service teachers.

CFA graphical and formulaic representation. CFA is commonly represented using a path diagram, which is a diagram using circles for the latent variable and squares or rectangles for the observed variables. These shapes are then connected using a series of specific lines and/or arrows. A single-headed arrow indicates a direction of assumed causal influence, and double-headed arrows represent covariance among latent variables. Another representation of a CFA model is the equation:

$$x = \lambda\xi + \delta$$

[4]

Where x is the vector of the observed variables i, λ (lambda) is the matrix of loadings connecting the latent variables ξi to the observed variables xi, ξ is the vector of common factors, and δ is the vector that represents the measurement error or it known as unique factors (Fox, 1983). At this time, the path diagram and equation for the CFA model on the modified CALI have yet to be determined. The structure and hypothesized factors are contingent upon completion of the first phase of this analysis.

Multiple Regression. Lastly, to further investigate the second research objective, Multiple Regression techniques were required. Multiple Regression is a statistical analysis that is a part of the family of analyses called the General Linear Model (GLM). The goal of the GLM is to explain the most amount of variance in the data (Hahs-Vaughn & Lomax, 2013). By explaining the most amount of variance, the GLM identifies when the model used explains the results better than chance (Stevens, 2009). The goal is to best explain the greatest amount of variance in the data possible. The statistical procedures of the GLM test for overall model significance, as well as the effect of any specific independent variable in the model. The overall model significance informs if the variables used in the model significantly predicted the outcome and the likelihood that these predictions were not due to chance. This is typically reported via the *F*-statistic for a number of GLM analyses, including Multiple Regression. Secondly, the GLM can test for the individual contribution of independent variables to the model. This is particularly important for regression models. Specifically, Multiple Regression measures the contribution of certain independent variables on the dependent variable (Keith, 2006). This analysis requires one or more categorical or continuous independent variables and one continuous dependent variable.

Prior to conducting a Multiple Regression Analysis, several statistical assumptions must be explored. The main statistical assumptions for Multiple Regression include: (1) Independence, (2) Normality, (3) Linearity, and (4) Homoscedasticity (Keith, 2006). The assumption of Independence posits that the variance in the variable is free, as opposed to the observed scores, or that the value of one observation does not influence or affect the value of other observations. The Durbin-Watson statistic can be used to identify if this assumption has been met (Hinkle, Wiersma, & Jurs, 2005). The assumption of Normality is tested for the normal distribution of residuals in the data across the variables (Shapiro & Wilk, 1965). Linearity measures the linear relationship between the dependent and independent variables. This is typically investigated using fit lines and scatterplot analyses. Lastly, Homoscedasticity tests for patterns across the entire line of fit (Keith). The data points should be free of any obvious patterns or clusters. This assumption is assessed in similar fashion to linearity. Other considerations for Multiple Regression must also be briefly addressed. The data must be investigated for the effects of multicollinearity, which occurs when two or more independent variables are highly correlated (Keith). This can affect the overall interpretation of the results and should be assed using VIF (variance inflation factor) and tolerance values as well as the correlation matrix. Lastly, Multiple Regression analyses are susceptible to biased effects from extreme outliers in the data. Using values like Cook's D and Mahalanobis Distance help detect significant outliers that may affect the results can be identified (Keith). The current study employed a specific subset of Multiple Regression analysis called Moderated Multiple Regression to investigate the relationship between assessment knowledge and performance and is discussed in more detail in the next section.

Moderated Multiple Regression. The second research objective of this study investigated the role of confidence in assessment knowledge as measured by the modified CALI and how this is related to the edTPA exam. This objective aimed to determine the relationship between confidence and assessment knowledge on the performance assessment outcomes as observed on the edTPA. The statistical approach for this objective was to conduct a Moderated Multiple Regression investigating the impact of confidence (i.e., the Moderator, or *M*) on assessment knowledge's (i.e., the main Independent Variable, or *X*) relationship to the outcome (i.e., the edTPA, which is the Dependent Variable, or *Y*). Moderation between *X* and *Y* occurs when the magnitude of the casual effect is influenced by at least one additional variable (Hayes, 2013). The term moderation and interaction are used commensurately in quantitative research.

Moderation Analysis answers the question of when (i.e., when does confidence impact edTPA performance). In other words, moderation is used to determine whether the size and sign of the effect of *X* on *Y* depends on the influence of (a) moderator variable(s). The effect of *X* on variable *Y* is moderated by *M* if its size, sign, or strength depends on or can be predicted by *M*. If these conditions are met, then *M* and *X* interact to influence Y. In the context of this research objective, M is confidence and the analysis will explore to what degree confidence strengthens or weakens the effect of classroom assessment knowledge (X) as measured by the modified CALI on performance assessment outcomes (Y, or the edTPA).

Moderation is used to determine the boundary condition for an association between two variables (Hayes, 2013). A boundary condition outlines the environment for when an association exists, or the direction of cause is known. Confidence will not necessarily produce a specific direction of performance or improvement. As Bandura's (1977) framework suggests, low confidence of self-efficacy can be produced from a low-perceived value in the task or be a factor of task difficulty. By exploring these boundary conditions, the analysis can answer the "when" questions such as under what circumstances, or for which types or people does X exert its effect on Y. Moderation Analysis is performed by testing the interaction between M and X in a model of Y. Testing an interaction occurs when a researcher quantifies and describes the bounded nature of the effect of X on Y at various values of the moderator (Hayes, 2013).

Moderated effect hypothesis. Figure 2 addresses the relationship between classroom assessment knowledge and edTPA performance and the hypothesized moderator confidence. The goal of the moderator analysis was to determine the following: (1) if confidence is a moderator of the relationship between the modified CALI and edTPA scores, and (2) if confidence is found to be a moderator, to what degree it demonstrates this moderator effect. The moderation model diagramed in Figure 2 can be expressed with the following equation:

$$Y = i_1 + b_1 X + b_2 \mathrm{M} + b_3 XM + e_y$$

[5]

Where X represents the outcome variable, M represents the moderator, and XM is the interaction between these two variables. The b_1 denotes the coefficient of the independent variable, b_2 the coefficient of the moderator, and b_3 the coefficient of the interaction term, which is a product of the independent variable and the moderator. The residuals in the equation are represented by e_y and the intercept of the equation is i_1.

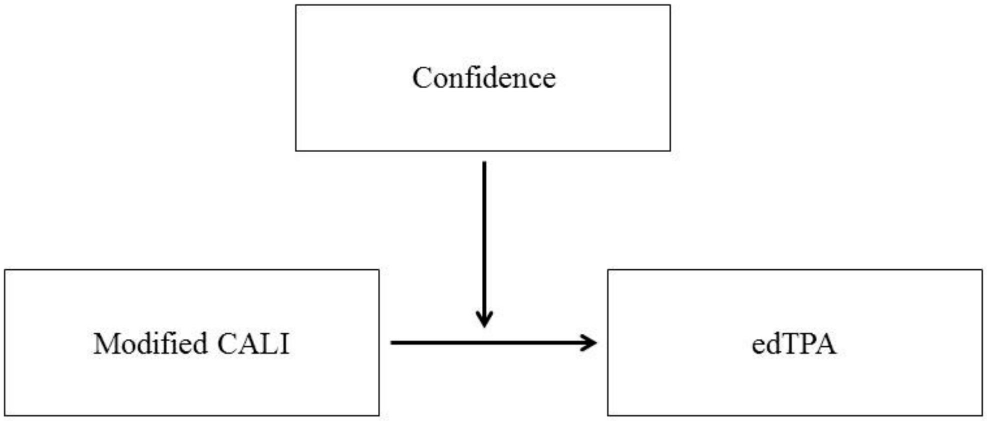


Figure 2. Confidence as a moderator of assessment knowledge and performance outcomes.

Since b_1 and b_2 are the coefficients for the modified CALI and confidence, respectively, they will need to be transformed for meaningful interpretation. If the coefficients were left untransformed, their value and tests of significance would have no substantive interpretation. In other words, by transforming these variables to a standardized scale, the relationship between the CALI score and participant confidence can be interpreted relative to each other and participant performance. Therefore, it is not logical to describe these coefficients without transformations. Variable mean centering was used to transform confidence and modified CALI scores prior to the analysis to increase the interpretability of the coefficients (Aiken, West, & Reno, 1991). Mean centering was accomplished by subtracting a constant, the mean, from every value in the data set. The resulting transformed values of, *M’* or *X’*, had a mean of zero and units measured in standard deviations. After mean centering, b_1 and b_2 were interpreted as the estimates when the other variable was conditioned to the mean of the *X’* or *M’*, respectively, instead of zero. In other words, b_1 estimates the difference in *Y* between two cases that differ by one unit on *X* among cases that were average on *M*. The inverse is true for b_2. Mean centering has no impact on the variance explained (R^2), the statistical significance of the overall model, or the inferences about the true b_3 (τb_3) coefficient. For all covariates, the value was set to their respective means (Hayes, 2013).

Visualizing moderation. To make the interpretability of moderation easier, a set of estimates of *Y* was generated from various combinations of *X* and *M*, using the non-centered mean regression model and the plotting $\hat{Y}$ as a function of *X* and *M*. The non-centered mean values of *X* and *M* were used as they are within the realm of plausible values for the measurement scales of the variables.

Probing an interaction. There is an inherent chance component to the estimate of *X’s* effect on *Y* at any chosen value of *M*. The likelihood of chance affecting this relationship is related to sampling error that occurs at each and every value of *M*. To accommodate this uncertainty, a set of post-interaction interferential tests were used to establish where, in the distribution of confidence, there was an effect on edTPA scores that was different from zero and where it does not. In other words, if the moderation effect reports significance, the degree of the moderation was explored through additional analyses. In order to probe for the effects of the interaction, a simple slopes analysis can be conducted (Aiken, West, & Reno, 1991; Rogosa, 1981). This test would report differences in the relationships between the predictor and the outcome variables according to the moderator at both high and low levels of the moderator. These high and low values are typically defined by standard deviation values above and below the mean. Then, the slopes representing the moderator at these high and low levels can be examined for significance. Interaction effects were investigated between the various subcomponent scores and edTPA outcome scores – edTPA Total score and edTPA Assessment scores. This indicates when confidence impacts performance across the distribution of scores, and allows for identifying the range of scores that are impacted by confidence.

Summary

This chapter presented the methodological approach used in the current study. In order to accomplish the two objectives, this study required both Rasch Analyses, as well as analyses that explore and evidence the underlying structure within the assessment knowledge and assessment confidence constructs in the modified CALI. Additionally, the hypothesized moderating impact of assessment confidence between the assessment knowledge and performance-based outcome relationship (i.e., the edTPA) was analyzed using Moderated Multiple Regression. This chapter also outlined the context and procedures of this study. In the following two chapters, the Results (Chapter 4) and Discussion (Chapter 5) are presented.

CHAPTER IV

RESULTS

The first objective of this study was to investigate the psychometric properties (i.e., reliability and validity) of the modified CALI, developed from the original CALI (Mertler, 2003). Specifically, this study examined assessment knowledge and understanding (i.e., assessment literacy) within a sample of pre-service teachers by investigating the psychometric properties (i.e., content and construct validity, internal consistency reliability) of participant scores on the CALI, as well as a measure of confidence after each question on the CALI (i.e., assessment confidence). Thus, this study evaluated assessment literacy in pre-service teachers and their assessment knowledge confidence. Research Question 1 and 1A state:

1. What are the psychometric properties of the newly-developed assessment literacy and confidence measure for pre-service teachers?

   1A. What is the internal structure (i.e., unidimensional or multidimensional) of the modified CALI?

The second objective of this study was to investigate the impact of assessment literacy and assessment confidence on performance assessment scores. All pre-service teacher education students within the target population were required to take the edTPA performance assessment to graduate. The edTPA is designed to evaluate pre-service teacher readiness across the domains of Planning, Instruction, and Assessment producing several sub-scale scores from fifteen rubrics. The second research question in this study evaluated the relationship between assessment knowledge (i.e., derived from the CALI), assessment confidence, and high-stakes performance assessment outcomes:

1. What is the impact of assessment confidence on the relationship between pre-service teachers' assessment literacy and performance assessment scores?

Research Question 1

The first research question asked: "What are the psychometric properties of the newly-developed assessment literacy and confidence measure (i.e., the modified CALI) for pre-service teachers?" In order to investigate this question, Rasch Analysis was conducted. The Rasch Analysis results (i.e., the psychometric properties of the measure) are presented below following the demographic and basic CALI descriptive statistics. Additionally, Research Question 1A examined the internal structure of the items on the modified CALI. These results are also presented in the subsequent sections.

Descriptives

The sample from the initial pilot study (*N* = 165) contained 45 males (27.3%) and 120 females (72.7%). This is consistent with current undergraduate enrollment trends as well as the general population of teachers across the nation (Peter, Horn, & Carroll, 2005). Of these participants, the average age was 21.04 (*SD* = 1.66; *Mdn* = 21, IQR = 1), with a range of 19 to 29 years old. The majority of participants were White/Caucasian (*n* = 157, 92.5%), with 4.8% of the sample reporting other races (e.g., African-American/Black, Hispanic, Asian). There were 43 (26.1%) first-generation college students in this sample. Additionally, the average self-reported cumulative Grade Point Average (GPA) was 3.52 (*SD* = .39; *Mdn* = 3.52, IQR = .50). Forty-two-point four percent of the participants were in an Early Childhood Education (ECED) program, 25.5% were in Middle Childhood Education (MCED), 27.3% were in Adolescent Education (AYA), and 4.8% were in Special Education (SPED).

Year in school is established according to how many credit hours the student has received and does not indicate how many he/she has completed within their teacher education program. For example, it is likely that some of the 4th- and 5th-year students may have changed majors. Based on these university credit hour requirements, there were 32 students in their second year (19.5%), 73 in their 3rd-year (44.5%), and 59 were 4th-year students (36.9%) or higher ($N = 164$). The highest level of education obtained by the students' parents or legal guardians was also recorded as an approximation of socioeconomic status (Sirin, 2005). For maternal education level ($N = 163$), 25.8% of students indicated that their mothers received a High School Diploma or General Education Diploma (GED) and 32.5% received a Bachelor's Degree. For paternal education ($N = 163$), similar to maternal, 31.3% of the participants' fathers received a High School Diploma or GED, and 29.4% received a Bachelor's Degree.

All but two participants in this sample reported some form of classroom experience ranging from observations to teaching. Specifically, 92.1% ($n = 152$) completed classroom observations, 39.4% ($n = 69$) performed teacher assistant duties, and 46.1% ($n = 76$) reported student teaching (i.e., supervised teaching). The vast majority of participants (90.3%) indicated feeling "Somewhat Prepared" ($n = 79$, 47.9%) or "Very Prepared" ($n = 70$, 42.4%) to be a teacher based on their training. However, pre-service teachers in this sample felt slightly less prepared to assess student learning (i.e., compared to feeling prepared to be a teacher) with 1.2% (*n* = 2) of the participants reporting being “Very Unprepared,” 21.8% (*n* = 36) indicating “Somewhat Unprepared,” 61.2% (*n* = 101) being “Somewhat Prepared,” and 15.8% (*n* = 26) feeling “Very Prepared.”

Students responded to other survey items pertaining to their assessment-related coursework and workshop participation and their self-perceptions of preparedness to be a teacher. Seventeen-point six percent (*n* = 29) of the sample reported taking a course solely focused on assessment, with 8.4% (*n* = 14) having completed two or more assessment-only courses. In this sample (*N* = 161), 7.5% (*n* = 12) indicated attending a workshop with an assessment-only focus, and 1.2% (*n* = 2) attended more than one assessment-only workshop. Courses that had at least one assessment component/unit, but were not solely focused on assessment, were attended by 71.4% (*n* = 115) of the participants (*N* = 161), with 68.5% (*n* = 113) attending more than one. Workshops with an assessment component/unit were attended by 14.3% (*n* = 23) of the participants, and 7.2% (*n* = 12) attended more than one.

Finally, the pilot sample’s average CALI score was 18.36 (*SD* = 3.71) out of a possible 35 multiple-choice (i.e., dichotomously scored) questions. This indicates that participants correctly responded to 52.5% of the CALI questions on average. The pilot sample’s average confidence score was 2.55 (*SD* = .51), which was calculated from the 35-item total score and divided by the number of items for an average on the Likert scale (i.e., 0 = Completely Unconfident to 4 = Completely Confident). The sample’s average confidence of 2.55 means that they had a slightly above neutral amount of confidence (i.e., between “Neither Confident nor Unconfident” and “Mostly Confident”) in their knowledge of assessment. Table 2 (see below) summarizes the pilot sample demographic and other descriptive variables from the preceding paragraphs. Total CALI and Assessment Confidence scores were used for each categorical demographic and descriptive variable in order to provide some evidence of group equivalence in this measure development pilot sample. Continuous variable descriptive statistics including age and GPA were included in the table as well.

Table 2

Pilot Sample Variable Descriptive Statistics: Classroom Assessment Literacy Inventory (CALI) Total Scores and Assessment Confidence Total Scores (N = 165)

Variable		CALI Total			Assessment Confidence Total		
	n	*M/Mdn*	*SD*/IQR	Min/Max	*M/Mdn*	*SD*/IQR	Min/Max
Gender							
Male	45	18.67	3.33	11/27	2.70	.47	1.49/3.71
Female	120	18.24/19.00*	3.85/5	4/27	2.49/2.52*	.51/.60	.63/3.57
Age	165	21.04/21.00*	1.66/1.00	19/29	--	--	--
Race							
White/Caucasian	157	18.50/19.00*	3.64/4.00	4/27	2.55/2.60*	.49/.57	.63/3.71
Other	8	15.63	4.34	10/22	2.51	.77	1.00/3.49
1st Generation College Student							
Yes	43	18.26	3.95	5/27	2.60	.52	1.14/3.49
No	122	18.39	3.64	4/27	2.53/2.60*	.50/.58	.63/3.71
GPA	165	3.52/3.52*	.39/.50	1.00/4.00	--	--	--
Program							
ECED	70	18.36/19.00*	4.05/4.00	4/27	2.37/2.39*	.55/.64	.63/3.37
MCED	42	18.40	3.36	11/24	2.74	.37	2.17/3.57
AYA	45	18.09	3.72	10/27	2.62	.49	1.49/3.71
Other	8	19.63	2.56	14/22	2.68/2.69*	.35/.44	2.31/3.37
Year							
Sophomore	32	17.94	4.29	5/24	2.37	.66	.97/3.57
Junior	73	18.68	3.85	4/27	2.53	.44	.63/3.57
Senior or Higher	59	18.12	3.19	10/24	2.54	.47	1.49/3.71
Mother's Education							
HSD/GED or Less	42	18.36	4.08	5/27	2.51	.63	.63/3.37
Some College/Associate/Tech	38	18.50	3.10	11/23	2.66	.35	1.80/3.49
Bachelor's Degree	53	18.17/19.00*	4.03/5.00	4/24	2.52	.45	1.60/3.57

Master’s/Doctoral/Professional	30	18.37	3.54	10/24	2.45	.56	.97/3.71
Father’s Education							
HSD/GED or Less	51	18.04/18.00*	4.21/4.00	4/27	2.58	.50	1.14/3.71
Some College/Associate/Tech	42	18.76	3.33	12/27	2.52	.53	.63/3.57
Bachelor’s Degree	48	18.19	3.75	10/24	2.52	.55	.97/3.49
Master’s/Doctoral/Professional	22	18.50	3.32	10/24	2.53	.33	1.77/3.09
Course(s) with Assessment							
Yes Took Course(s)	115	18.43	3.59	5/27	2.61	.47	1.00/3.71
No Did Not	46	17.96	4.08	4/27	2.36/2.41*	.54/.48	.63/3.49
Assessment-Specific Course							
Yes Took Course	29	17.38	3.26	10/23	2.51	.37	1.49/3.37
No Did Not	136	18.57/19.00*	3.78/4.00	4/27	2.55/2.60*	.53/.66	.63/3.71
Student Teaching Experience							
Yes	76	18.07/18.00*	3.79/5.00	4/24	2.61	.44	1.60/3.71
No	89	18.61	3.65	10/27	2.49/2.60*	.55/.54	.63/3.57
CALI Total	165	18.36/19.00*	3.71/5.00	4/27	--	--	--
Assessment Confidence Total	165	--	--	--	2.55/2.60*	.51/.59	.63/3.71

Note. Groups and continuous variables denoted with asterisks next to the values in the *M/Mdn* columns indicate non-normal distributions.

Relationships between these demographic variables and the Total CALI and Assessment Confidence scores were examined to provide evidence of group equivalence (see Table 3). The measurement levels of the variables and the skewness of CALI Total and Confidence scores dictated the statistical tests selected and if the analyses were parametric or nonparametric. Pearson or Spearman correlations were used to examine relationships between two continuous or ordinal variables. Independent *t*-Tests and One-Way ANOVAs (i.e., or their nonparametric equivalents) were selected to investigate CALI Total and Confidence score differences between groups with two or three or more levels of the categorical variable, respectively. The variables that had a statistically significant relationship with the Total CALI scores were race (p = .048) and GPA (p = .044). Several variables reported a significant relationship with Assessment Confidence scores including gender (p = .015), age (p = .006), program (p = .003), year in school (p = .032), and having taken a course with an assessment component (p = .005).

Table 3

Relationships between Pilot Sample Variables and Classroom Assessment Literacy Inventory (CALI) Total Scores and Assessment Confidence Total (Average) Scores (N = 165)

Variable	CALI Total		Assessment Confidence Total	
	Statistical Test	p	Statistical Test	p
Gender	U = 2645.50, Z = -.200	.841	U =2037.00, Z = -2.427	.015*
Age	r_s = -.014	.858	r_s = .214	.006**
Race	U = 368.50, Z = -1.978	.048*	U = 617.50, Z = -.080	.936
1st Generation College Student	t(163) = .208	.835	U =2400.00, Z = -.828	.408
GPA	r_s = .157	.044*	r_s = .002	.980
Program	H(3) = 1.829	.609	H(3) = 13.940	.003**
Year	F(2,161) = .604	.548	F(2,161) = .6043.504	.032*
Mother's Education	H(3) = .117	.990	F(3,159) = 1.190	.315

Father's Education	$H(3) = .433$	.933	$F(3,159) = .163$	.921
Course(s) with Assessment	$t(159) = -.733$	.465	$U = 1896.00, Z = -$	$.005^{**}$
Assessment-Specific Course	$U = 1562.00, Z = -$	.078	2.804	.622
Student Teaching	1.761	.458	$U = 1857.00, Z = -.493$	.504
Experience	$U = 3156.00, Z = -.742$	.058	$U = 3177.50, Z = -.669$	--
Assessment Confidence	$r_s = .148$		--	
Total				

Note. $^{*}p < .05$, $^{**}p <.01$, $^{***}p < .001$.

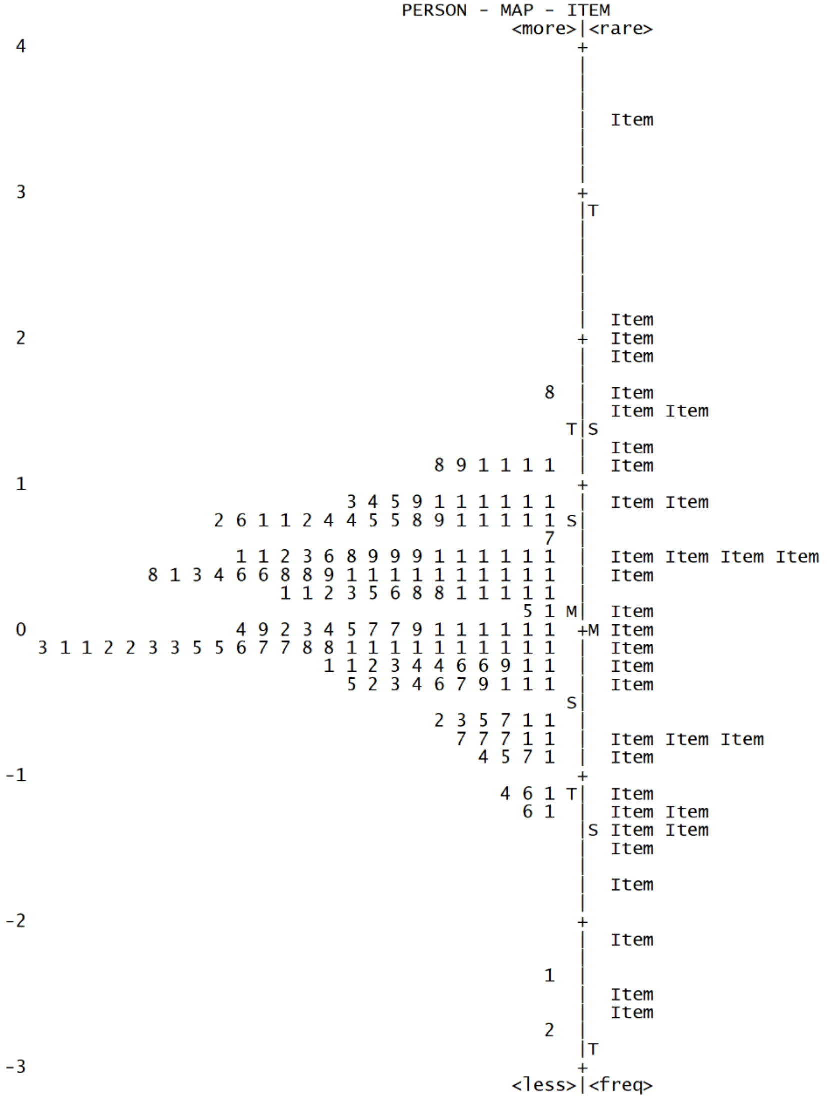
Rasch Analysis

The first Rasch Analysis on the CALI scores used the pilot sample data, which included responses to all 35 multiple-choice content questions and 35 confidence scale questions. No questions or persons were removed from this initial analysis. After the presentation of the initial pilot results in the first Rasch Analysis, measurement refinement was conducted, or the elimination of poorly fitting items. This process then led to a second Rasch Analysis of the now modified CALI (i.e., after item removal). Both of these analyses were conducted on responses from the same pilot sample. Thus, the following sections of results will contain: (1) the pilot sample 35-item CALI Rasch Analysis, (2) the pilot sample 35-item Assessment Confidence Rasch Analysis, (3) the pilot sample 25-item CALI Rasch Analysis, (4) the pilot sample 25-item Assessment Confidence Measure Rasch Analysis, (5) the pilot sample 25-item CALI Rasch Principal Components Analysis (PCA), (6) the pilot sample 12-item CALI Component 1 Rasch Analysis, (7) the 12-item CALI Component 1 Assessment Confidence Rasch Analysis, (8) the pilot sample 13-item CALI Component 2 Rasch Analysis, and (9) the pilot sample 13-item CALI Component 2 Assessment Confidence Rasch Analysis.

Pilot CALI Rasch analysis. The Rasch Dichotomous Model was used to analyze these data. All 35 multiple-choice items were scored as "Correct" (Coded 1) or

"Incorrect" (Coded 0). The initial analysis converged after three iterations. The average score across all participants was 18.30 points or 52.3% correct responses (*SD* = 3.60). Initial investigation of the item-person map (Figure 3) revealed the range of items and persons to be between -2.5 and approximately 3.5 logits. Specifically, items fell between -2.5 and 3.5 logits and persons appeared between -2.75 and 1.75 logits. Items and persons were largely contained between -1.50 and 1.50 logits; however, some items fell outside this range. Four items (items 31, 28, 30, and 7) were located above the highest performing person (1.75 logits), indicating that these items were too difficult for the sample. Alternatively, three items (items 9, 15, and 1) were easily answered by all respondents and were located below -2.0 logits. Several sets of items were located at the same difficulty level on the vertical ruler, which may indicate redundant content. For ability, two persons were located below -2.0 logits, and one person was above 1.5 logits.

Item summary statistics revealed relative congruence and minimal misfit data between the real and model values. Therefore, only the real item and person summary statistics are reported. Item Root-Mean-Square-Error (RMSE) was low (.21, *SD* = 1.41). Additionally, item separation (6.78) was high and item reliability (.98) was strong, suggesting equal distribution in items across difficulty levels and consistency in item placement. High item separation (i.e., > 3 logits) coupled with high item reliability (i.e., > .9) implies that the person sample is large enough to confirm the item difficulty hierarchy and provide evidence of construct validity (Linacre, 1991). Items spanned the range of -2.61 to 3.54 logits with the expected mean of zero. Reviewing the summarized item infit and outfit statistics indicated a high infit maximum value (MNSQ = 1.20, ZSTD = 2.20) and outfit maximum (MNSQ = 1.70, ZSTD = 3.6).



```
                                    PERSON - MAP - ITEM
                                              <more>|<rare>
 4                                                  +
                                                    |
                                                    |
                                                    |
                                                    |  Item
                                                    |
                                                    |
                                                    |
 3                                                  +
                                                    |T
                                                    |
                                                    |
                                                    |
                                                    |
                                                    |
                                                    |  Item
 2                                                  +  Item
                                                    |  Item
                                                    |
                                                 8  |  Item
                                                    |  Item Item
                                                   T|S
                                                    |  Item
                                       8 9 1 1 1 1  |  Item
 1                                                  +
                               3 4 5 9 1 1 1 1 1 1  |  Item Item
                   2 6 1 1 2 4 4 5 5 8 9 1 1 1 1 1 S|
                                                 7  |
                     1 1 2 3 6 8 9 9 9 1 1 1 1 1 1  |  Item Item Item Item
             8 1 3 4 6 6 8 8 9 1 1 1 1 1 1 1 1 1 1  |  Item
                         1 1 2 3 5 6 8 8 1 1 1 1 1  |
                                               5 1 M|  Item
 0                   4 9 2 3 4 5 7 7 9 1 1 1 1 1 1  +M Item
   3 1 1 2 2 3 3 5 5 6 7 7 8 8 1 1 1 1 1 1 1 1 1 1  |  Item
                             1 1 2 3 4 4 6 6 9 1 1  |  Item
                               5 2 3 4 6 7 9 1 1 1  |  Item
                                                   S|
                                       2 3 5 7 1 1  |
                                         7 7 7 1 1  |  Item Item Item
                                           4 5 7 1  |  Item
-1                                                  +
                                             4 6 1 T|  Item
                                               6 1  |  Item Item
                                                    |S Item Item
                                                    |  Item
                                                    |
                                                    |  Item
                                                    |
-2                                                  +
                                                    |  Item
                                                    |
                                                 1  |
                                                    |  Item
                                                    |  Item
                                                 2  |
                                                    |T
-3                                                  +
                                              <less>|<freq>
```

Figure 3. Pilot Sample 35-Item Item-Person Map (*N* = 165). This vertical scale illustrates the placement of persons and items on a continuum representing the latent variable (i.e., assessment literacy). The left side of the scale presents the logit values, which are a common interval scale created by the Rasch Model with a mean of 0.0. Items are located on the right side of the scale. Items higher on the continuum were more difficult for persons in this sample to endorse correctly, while items lower on the continuum were easier. Persons are placed on the left side of the scale according to their ability and the degree of difficulty of the items.

Much like the items, person summary statistics revealed congruence and minimal misfit between the real and model values. The RMSE was .43 (*SD* = .45) with the person mean (.11) only slightly higher than the item mean of zero. Thus, the positive mean provided evidence of the presence of classroom assessment literacy within this sample. Person separation (1.04) and reliability (.52) were acceptable to low, with the person range of -2.70 to 1.66 logits being slightly restricted. Person infit and outfit statistics also revealed high infit maximum values (MNSQ = 2.12, ZSTD = 4.5) and outfit maximum values (MNSQ = 7.80, ZSTD = 4.30). The person raw score reliability (KR-20) was acceptable at .53, indicating moderate reliability in replicating participants' scores across survey administrations (Linacre, 2010).

Table 4

Pilot Sample 35-Item CALI Summary of Person Statistics (N = 165)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	18.4	35.0	-3.63	.41	1.00/-.1	1.08/-.1
P.SD	3.7	.0	.62	.02	.30/1.3	.80/1.2
S.SD	3.7	.0	.63	.02	.30/1.3	.80/1.2
Max	27.0	35.0	-2.13	.57	2.97/5.4	7.54/4.5
Min	4.0	35.0	-6.43	.40	.54/-2.9	.43/-2.3

Note. Real/Model RMSE = .43/.41, Real/Model True SD = .45/.47; Real/Model Separation = 1.06/1.16; Real/Model Person Reliability = .53/.57; Standard Error of Person Mean = .05; Coefficient Alpha (KR-20) = .54, SEM = 2.51.

Table 5

Pilot Sample 35-Item CALI Summary of Item Statistics (N = 35)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	86.5	165.0	.00	.20	.98/-.1	1.08/.5
P.SD	42.6	.0	1.41	.05	.10/1.0	.25/1.5
S.SD	43.3	.0	1.43	.05	.10/1.0	.26/1.5
Max	153.0	165.0	3.54	.42	1.20/2.2	1.70/3.6
Min	6.0	165.0	-2.49	.16	.71/-1.7	.59/-1.8

Note. Real/Model RMSE = .21/.20; Real/Model True SD = 1.39/1.39; Real/Model Separation = 6.76/6.85; Real/Model Item Reliability = .98/.98; Standard Error of Item Mean = .24.

Individual item misfit was analyzed to further investigate the reported high infit and outfit values. This analysis revealed a series of misfit items with infit or outfit values whose MNSQ exceeded 1.2 (Wright, Linacre, Gustafson, & Martin-Lof, 1994) or ZSTD value exceeding 2 (Wright & Masters, 1982). These items included Item 20, Item 5, Item 7, Item 30, Item 28, Item 21, and Item 31 (see Table 6). All other items had infit and outfit MNSQ values close to 1, and thus, were not misfitting.

Table 6

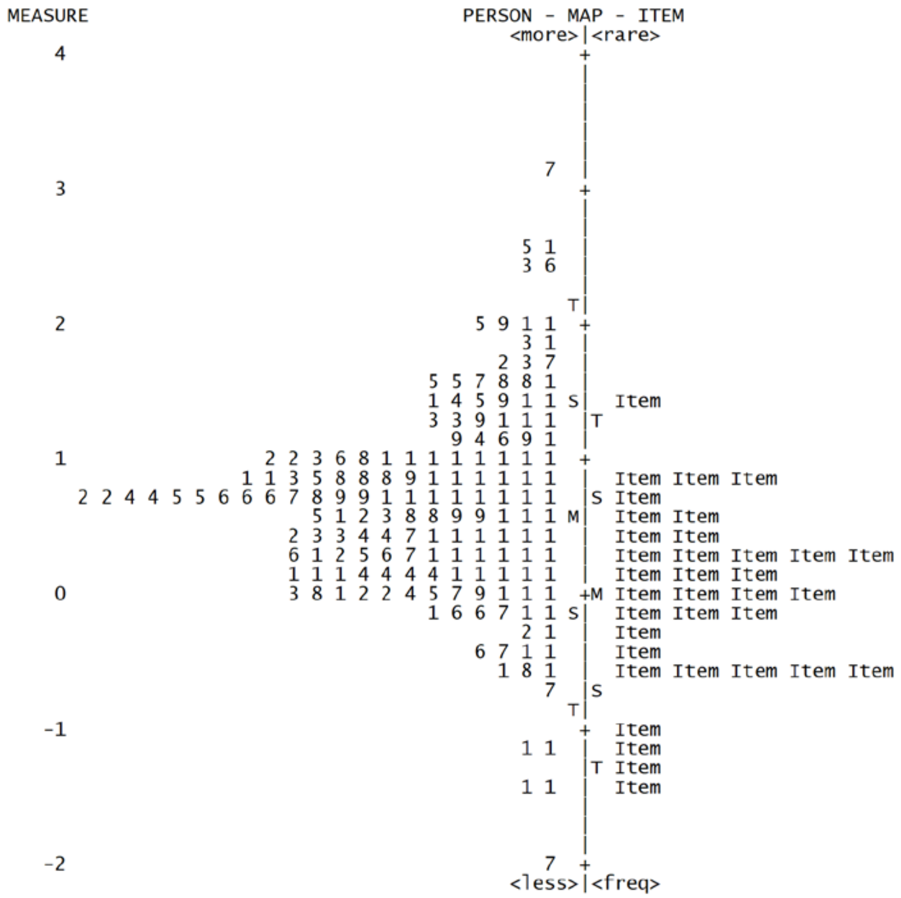
Pilot Sample 35-Item CALI Item Misfit Statistics (N = 35)

Item	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD	Point-Measure Correlation
31	2.00	.22	1.04/.3	1.41/2.6	85.5 (85.4)
21	1.52	.20	1.06/.6	1.26/2.3	80.6 (78.8)
28	3.54	.42	1.04/.2	1.24/2.4	96.4 (96.4)
30	2.10	.23	1.00/.0	1.17/1.4	86.7 (86.7)
7	1.90	.22	1.09/.7	1.15/2.3	84.2 (84.2)
5	1.44	.19	1.03/.3	1.09/1.5	78.8 (77.7)
6	-.94	.18	1.20/2.2	1.09/.7	69.7 (68.5)
20	.92	.17	1.16/2.3	1.09/.8	67.9 (68.5)
22	1.27	.18	.98/-.2	1.09/1.4	74.5 (74.7)
2	.45	.16	1.07/1.5	1.08/1.3	61.2 (62.2)
4	-.26	.16	1.07/1.3	1.07/1.0	58.8 (63.5)

14	1.48	.19	1.04/.4	1.09/.7	77.0 (78.2)
19	1.11	.18	1.07/.9	1.09/.8	71.5 (71.8)
27	.48	.16	1.03/.6	1.09/1.4	59.4 (62.4)
17	.37	.16	.98/-.4	1.08/1.3	64.2 (61.5)
18	.53	.16	1.08/1.6	1.07/1.0	54.5 (62.8)
10	.83	.17	1.00/.0	1.04/.5	68.5 (67.0)
8	-.72	.17	1.00/.0	1.02/.3	69.7 (70.2)
12	-.75	.17	.99/-.2	1.02/.2	70.3 (70.6)
26	.45	.16	1.00/-.1	1.01/.1	64.8 (62.2)
32	-1.29	.20	.96/-.3	.98/-.4	80.0 (79.0)
13	.18	.16	.98/-.4	.93/-.7	60.6 (61.1)
25	-1.07	.19	.93/-.7	.93/-.3	79.4 (75.7)
3	-1.36	.20	.96/-.3	.91/-.6	81.2 (80.2)
35	-.10	.16	.96/-.9	.95/-.9	65.5 (61.8)
29	-.05	.16	.93/-1.5	.92/-1.4	69.1 (61.6)
1	-2.08	.24	.92/-.4	.89/-.4	89.1 (88.3)
16	-1.25	.19	.89/-1.0	.90/-.8	80.6 (78.5)
24	-.40	.17	.90/-1.7	.88/-1.7	73.3 (65.4)
11	-1.76	.22	.89/-.7	.81/-1.0	86.7 (85.1)
33	-.75	.17	.89/-1.4	.88/-1.3	75.5 (70.6)
23	-1.49	.21	.86/-1.1	.86/-.9	83.0 (81.9)
34	-1.40	.20	.86/-1.2	.83/-1.1	81.8 (80.7)
9	-2.49	.28	.73/-1.2	.66/-1.4	93.3 (91.1)
15	-2.41	.27	.71/-1.4	.59/-1.8	92.7 (90.7)

Lastly, item point-measure correlations did not depart from their expected values and all correlations were positive. This correlation is similar to a point-biserial correlation, but involves the logit structure of the Rasch Model. Point-measure correlations are correlations between the observations of an item and the item real scores. These are crucial for evaluating if higher observations of the desired trait (i.e., real responses) correspond with an increased level of the latent variable. Negative or zero point-measure correlations indicate items or persons with response strings that contradict the variable. Conversely, when the correlation is high and exceeds its expected value, the item overfits or fits the Rasch model too perfectly

Pilot confidence Rasch analysis. The Rasch Rating Scale Model (RSM) was used to analyze these data as all 35 confidence items required a response on a 5-point Likert-scale, ranging from “Complete Unconfident” (Coded 0) to “Completely Confident” (Coded 4). The analysis converged after thirteen iterations. Initial investigation of the item-person map (Figure 4) revealed a large cluster of items between -1.5 and approximately 1.5 logits. This positioning was consistent for participants.



```
MEASURE                                PERSON - MAP - ITEM
                                           <more>|<rare>
    4                                            +
                                                 |
                                                 |
                                                 |
                                                 |
                                                 |
                                              7  |
    3                                            +
                                                 |
                                                 |
                                            5 1  |
                                            3 6  |
                                                 |
                                                T|
    2                                   5 9 1 1  +
                                            3 1  |
                                          2 3 7  |
                                    5 5 7 8 8 1  |
                                    1 4 5 9 1 1 S|  Item
                                    3 3 9 1 1 1  |T
                                      9 4 6 9 1  |
    1                 2 2 3 6 8 1 1 1 1 1 1 1 1  +
                    1 1 3 5 8 8 8 9 1 1 1 1 1 1  |  Item Item Item
      2 2 4 4 5 5 6 6 6 7 8 9 9 1 1 1 1 1 1 1 1  |S Item
                          5 1 2 3 8 8 9 9 1 1 1 M|  Item Item
                        2 3 3 4 4 7 1 1 1 1 1 1  |  Item Item
                        6 1 2 5 6 7 1 1 1 1 1 1  |  Item Item Item Item Item
                        1 1 1 4 4 4 4 1 1 1 1 1  |  Item Item Item
    0                   3 8 1 2 2 4 5 7 9 1 1 1  +M Item Item Item Item
                                    1 6 6 7 1 1 S|  Item Item Item
                                            2 1  |  Item
                                        6 7 1 1  |  Item
                                          1 8 1  |  Item Item Item Item Item
                                              7  |S
                                                T|
   -1                                            +  Item
                                            1 1  |  Item
                                                 |T Item
                                            1 1  |  Item
                                                 |
                                                 |
                                                 |
   -2                                         7  +
                                           <less>|<freq>
```

Figure 4. Pilot Sample 35-Item Confidence Item-Person Map (*N* = 165). This vertical scale illustrates the placement of persons and items on a continuum representing the latent variable (i.e., assessment confidence). The left side of the scale presents the logit

values, which are a common interval scale created by the Rasch Model with a mean of 0.0. Items are located on the right side of the scale. Items higher on the continuum were more difficult for persons in this sample to endorse, while items lower on the continuum were easier. Persons are placed on the left side of the scale according to their ability and the degree of difficulty of the items.

Item summary statistics revealed relative congruence and minimal misfit data between the real and model values. Therefore, only the real item and person summary statistics are reported. RMSE was low (.10, *SD* = .64). Additionally, item separation (6.21) was high and item reliability (.97) was strong, suggesting equal distribution in items across difficulty levels and consistency in item placement. High item separation (i.e., > 3 logits) coupled with high item reliability (i.e., > .9) implies that the person sample is large enough to confirm the item difficulty hierarchy and provide evidence of construct validity (Linacre, 1991). Items spanned the range of -1.44 to 1.45 logits with the expected mean of zero. Reviewing the summarized item infit and outfit statistics indicated a high infit maximum value (MNSQ = 1.86, ZSTD = 6.1) and outfit maximum (MNSQ = 1.77, ZSTD = 5.7).

Much like items, person summary statistics revealed congruence and minimal misfit between the real and model values. The RMSE was .23 (*SD* = .73) with higher person mean (.64) indicating the cluster observed in the item-person map. Thus, the positive mean provided evidence of the presence of classroom assessment literacy within this sample. Person separation (3.10) and reliability (.91) were acceptable; with a person range of -1.95 to 3.13 logits being slightly restricted. Person infit and outfit statistics also revealed high infit maximum (MNSQ = 2.45, ZSTD = 4.8) and outfit maximum (MNSQ = 3.52, ZSTD = 5.4). The person raw score reliability (Coefficient Alpha) was strong at

.92, indicating reliability in replicating participants' scores across survey administrations (Linacre, 2010).

Table 7

Pilot Sample 35-Item Confidence Measure Summary of Person Statistics (N = 165)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	89.2	35.0	.64	.21	1.02/-.2	1.03/-.1
P.SD	17.7	.0	.76	.02	.50/2.1	.52/2.0
S.SD	17.7	.0	.77	.02	.51/2.1	.53/2.0
Max	130.0	35.0	3.13	.35	2.46/4.8	3.52/5.4
Min	22.0	35.0	-1.95	.18	.26/-4.3	.26/-4.4

Note. Real/Model RMSE = .23/.21; Real/Model True SD = .73/.73; Real/Model Separation = 3.10/3.46; Real/Model Person Reliability = .91/.92; Standard Error of Person Mean = .05; Coefficient Alpha (KR-20) = .92, SEM = 4.90.

Table 8

Pilot Sample 35-Item Confidence Measure Summary of Item Statistics (N = 35)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	420.5	165.0	.00	.10	1.03/.1	1.03/.1
P.SD	69.7	.0	.65	.01	.27/2.2	.27/2.2
S.SD	70.8	.0	.66	.01	.27/2.3	.27/2.2
Max	555.0	165.0	1.45	.12	1.86/6.1	1.77/5.7
Min	240.0	165.0	-1.44	.09	.68/-3.3	.69/-3.2

Note. Real/Model RMSE = .10/.10; Real/Model True SD = .64/.64; Real/Model Separation = 6.21/6.60; Real/Model Item Reliability = .97/.98; Standard Error of Item Mean = .11.

Individual item fit was analyzed to further investigate the summary reported infit and outfit MNSQ values. This analysis revealed a series of misfit items with MNSQ infit or outfit values greater than 1.30. This is the suggested value for fit by Smith, Schumacker, and Bush (1995) for a sample of less than 500. The potentially misfitting items having MNSQ infit and/or MNSQ outfit values exceeding 1.30 include: Item 32, Item 33, item 1, Item 34, and Item 15. Additionally, the MNSQ values for these items did not exceed 1.86 which is also under the 2/-2 value suggested by Smith (1992). All other items had infit and outfit MNSQ values close to 1. Lastly, item point-measure correlations did not depart from their expected values and all correlations were positive.

Table 9

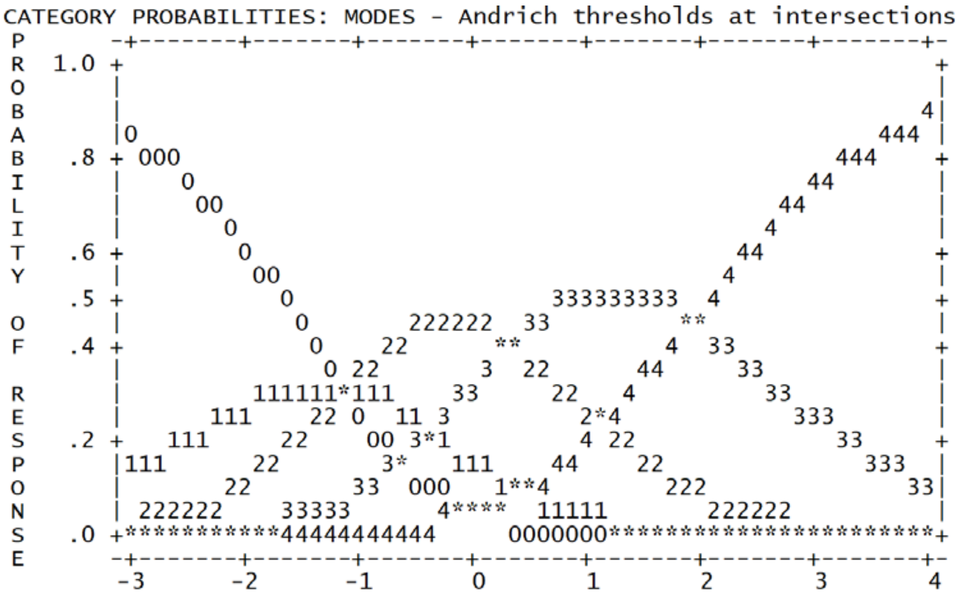
Pilot Sample 35-Item Confidence Measure Item Misfit Statistics (N = 35)

Item	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ ZSTD	Point-Measure Correlation
34	-.55	.10	1.86/6.1	1.77/5.7	.50 (.50)
1	-.52	.10	1.56/4.3	1.73/5.5	.34 (.50)
33	.02	.09	1.52/4.3	1.73/5.5	.52 (.52)
32	-.57	.10	1.46/3.6	1.41/3.3	.45 (.49)
15	-1.44	.12	1.38/3.0	1.16/1.3	.55 (.43)
31	1.45	.09	1.28/2.7	1.30/2.7	.33 (.55)
16	-.99	.11	1.21/1.8	1.28/2.3	.56 (.47)
2	-.02	.10	1.24/2.0	1.24/2.0	.42 (.52)
7	.31	.09	1.19/1.7	1.21/1.8	.42 (.54)
28	.84	.09	1.14/1.4	1.17/1.6	.48 (.55)
8	-.35	.10	1.06/.6	1.12/1.1	.53 (.51)
24	-.47	.10	1.06/.6	1.12/1.0	.57 (.50)
30	.55	.09	1.03/.3	1.06/.6	.55 (.54)
11	-1.18	.11	1.05/.5	1.01/.2	.44 (.45)
9	-1.35	.12	1.01/.1	.94/-.5	.49 (.44)
29	.04	.09	.98/-.1	.99/-.1	.62 (.53)
14	.84	.09	.97/-.3	.96/-.3	.46 (.55)
35	-.01	.10	.97/-.2	.95/-.4	.53 (.52)
17	.31	.09	.96/-.3	.92/-.7	.50 (.54)
12	-.50	.10	.94/-.5	.91/-.8	.57 (.50)
20	.49	.09	.89/-1.0	.91/-.8	.65 (.54)
4	.17	.10	.89/-1.0	.87/-1.1	.46 (.53)

25	-.16	.10	.83/-1.6	.88/-1.1	.58 (.52)
27	.44	.09	.84/-1.5	.87/-1.3	.57 (.54)
10	.20	.09	.82/-1.7	.86/-1.3	.58 (.53)
13	-.15	.10	.84/-1.5	.83/-1.6	.59 (.52)
18	.57	.09	.80/-2.0	.84/-1.5	.57 (.54)
3	-.17	.10	.82/-1.6	.80/-1.9	.50 (.52)
5	.14	.09	.80/-1.8	.81/-1.8	.53 (.53)
22	.71	.09	.81/-1.9	.81/-1.9	.55 (.55)
19	.35	.09	.75/-2.5	.80/-2.0	.58 (.54)
26	.90	.09	.77/-2.5	.79/-2.2	.54 (.55)
23	-.58	.10	.78/-2.1	.72/-2.8	.62 (.49)
21	.34	.09	.72/-2.8	.77/-2.3	.58 (.54)
6	.33	.09	.68/-3.3	.69/-3.2	.58 (.54)

Next, the summary of category structure and observed average of endorsements was investigated. The observed frequencies and percentages for each possible response category revealed higher endorsement for “Neither Confident Nor Unconfident” at 30% of all responses and “Mostly Unconfident” at 10% of responses. The “Completely Unconfident” category only accounted for 5% of responses, while “Mostly Confident” had 38% and “Complete Confident” had 18%. This pattern of endorsements supports the positive cluster of persons above the mean on the item-person map, indicating slightly above average assessment confidence.

Infit and outfit MNSQ values appeared normal and did not markedly exceed 1.30. Also, the Andrich thresholds (i.e., difficulty between endorsing each response option) displayed a monotonically increasing average. The greatest increase was between “Mostly Confident” (.33) and “Completely Confident” (1.93). The large logit gap between these categories demonstrated the relative ease in selecting between the “Mostly Confident” and “Completely Confident” categories compared to other adjacent categories. As shown in Figure 5 below, the lower three Likert categories have closer and less defined crests between “Completely Unconfident,” “Mostly Unconfident,” and “Neither Confident nor Unconfident.”



```
CATEGORY PROBABILITIES: MODES - Andrich thresholds at intersections
P      -+-------+-------+-------+-------+-------+-------+-------+-
R  1.0 +                                                         +
O      |                                                         |
B      |                                                        4|
A      |0                                                    444 |
B   .8 + 000                                              444    +
I      |    0                                           44       |
L      |     00                                       44         |
I      |       0                                     4           |
T   .6 +        0                                  44            +
Y      |         00                               4              |
    .5 +           0                  333333333  4               +
O      |            0       222222  33         **                |
F   .4 +             0    22      **          4  33              +
       |              0 22       3  22      44     33            |
R      |         111111*111    33     22   4         33          |
E      |      111    22 0  11 3         2*4            333       |
S   .2 +   111     22    00 3*1         4 22              33     +
P      |111      22       3*   111    44    22              333  |
O      |       22       33  000   1**4        222              33|
N      | 222222    33333      4****  11111       222222          |
S   .0 +***********4444444444      0000000***********************+
E      -+-------+-------+-------+-------+-------+-------+-------+-
       -3      -2      -1       0       1       2       3       4
```

Figure 5. Pilot Sample 35-Item Confidence Category Probabilities (*N* = 165). This figure indicates the probability of endorsement for each of the Likert scale confidence categories, starting with “Completely Unconfident” on the left and ending with “Complete Confident” on the right. The strength of the peak between each category indicates the discrimination between response levels.

Table 10

Pilot Sample 35-Item Confidence Measure Category Thresholds and Fit Table (N = 35)

Category Label	Observed Count	Observed Average	Sample Expected	Infit MNSQ	Outfit MNSQ	Andrich Threshold	Category Measure
0	284 (5%)	-.70	-.84	1.17	1.23	None	(-2.69)
1	556 (10%)	-.13	-.22	1.12	1.23	-1.19	-1.30
2	1704 (30%)	.23	.30	.91	.91	-1.08	-.20
3	2178 (38%)	.84	.87	1.00	.98	.33	.33
4	1053(18%)	1.66	1.56	.90	.92	1.93	(3.16)

Measure refinement. Based on the results of the above Rasch Analysis, changes were made to the measure for the second phase of this study. The initial pilot phase presented participants with all CALI content questions, as illustrated in the above analyses. Several items had aberrant fit statistics and were eliminated. A total of ten items were removed from the modified CALI due to item misfit. These misfit statistics revealed extreme MNSQ and ZSTD Infit and/or Outfit statistics, point-measure correlations, and content redundancies. These items included, Item 6, Item 7, Item 14, Item 18, Item 19, Item 20, Item 21, Item 27, Item 28, and Item 34. The procedures and justifications for removing these ten items are outlined below.

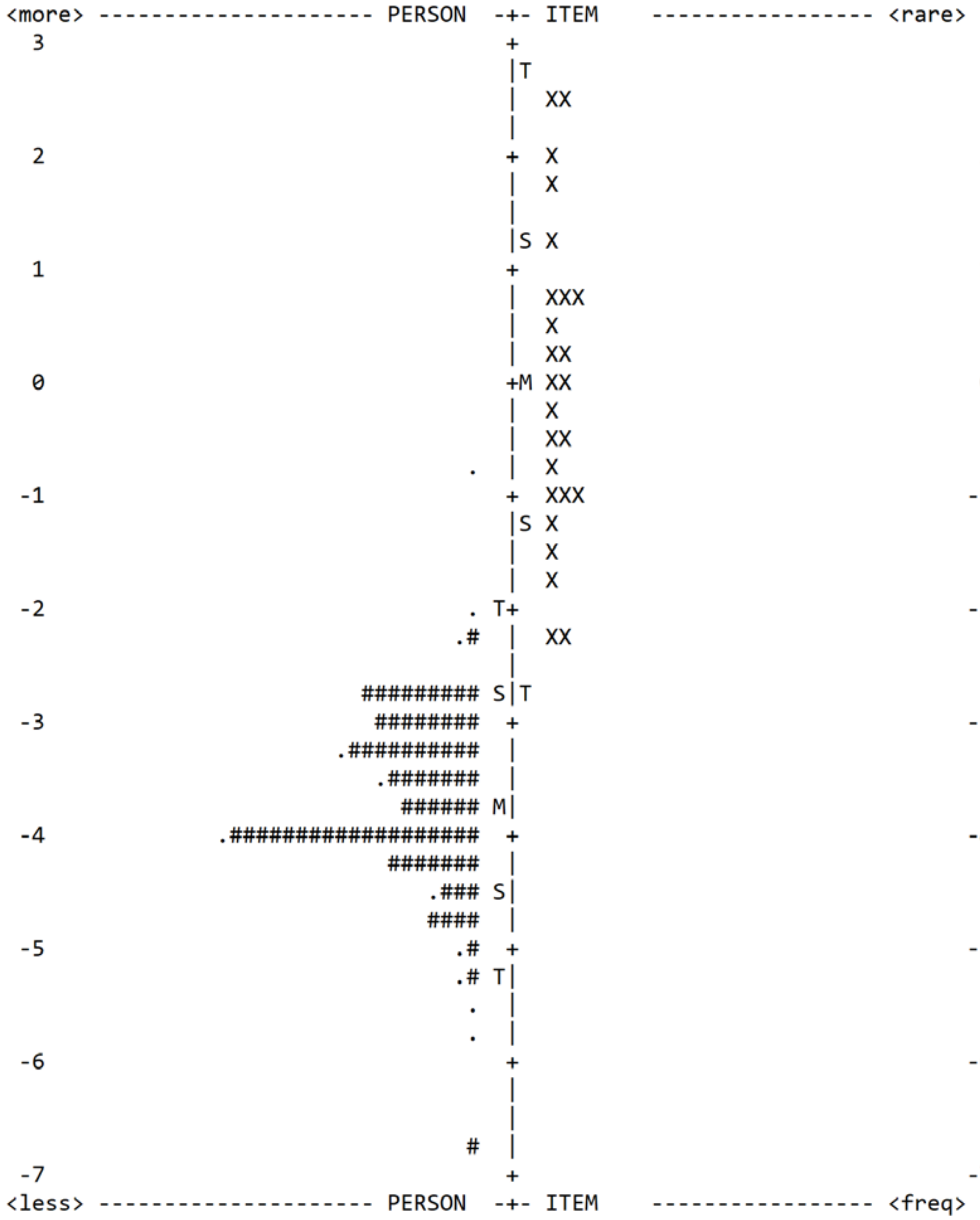
Misfitting items are those with response patterns which are too predictable or too unpredictable to meet the Rasch model probabilistic expectations (Boone, Staver & Yale, 2014). Misfitting items have infit and/or outfit MNSQ and/or ZSTD statistics that are extreme compared to commonly accepted guidelines (Boone et al., 2014). Infit violations occur when a response pattern is irregular, such as a person correctly endorsing complex questions but incorrectly endorsing any easy question. Outfit violations appear when response patterns are irregular or off-target (Andrich, 1988). For instance, outfit problems may surface when a person seems to randomly endorse both complex and easy items, but does not have a response pattern that clearly illustrates their ability/presence of the trait.

Both infit and outfit statistics are residual-based and designate the degree of misfit of observations to the Rasch model, with infit statistics using weighted squared residuals and outfit statistics using unweighted squared residuals that are summed and averaged (Bond & Fox, 2015). Items 28, 7, and 21 were removed due to MNSQ outfit values greater than 1.2 (Wright et al., 1994). Items 20 and 18 were removed due to high ZSTD infit and outfit values exceeding 2 (Wright & Masters, 1982). These fit values were investigated for both positive and negative ill-fitting values.

Additional information needed for investigating item performance are the point-measure correlations. Negative or zero point-measure correlations indicate items or persons with response strings that contradict the variable. Items 27 and 14 had low point-measure correlations. Conversely, when the correlation is high and exceeds its expected value, the item overfits or fits the Rasch model too perfectly. Items 19 and 6 reported low loadings on the contrast and high point-measure correlations. These aberrant fit statistics suggest the question may be measuring a different trait/construct or that the participants were not responding to the question using existing knowledge.

Lastly, Item 34 was removed due to content redundancies. The Rasch model displays the positions of all items on a continuum (i.e., the logit-based vertical ruler or item-person map) measuring the construct. When multiple items are located at the same logit difficulty on the continuum, the content of these items should be examined for redundancy. Item 34, specifically, was located at the same position on the vertical ruler with two other items. Upon further investigation, these items contained similar content knowledge at the same level of difficulty. Following the above modifications, the subsequent analysis of the CALI consisted of a 25-item version excluding the ten items removed as discussed above. The pilot sample's performance after removing these ten items is presented below.

25-Item modified CALI results. The results presented below investigated the performance of the pilot sample after removing ten items from the CALI. At this point, items were re-numbered out of 25. See Appendix D for numbering references. The total group average score was 59.2% correct responses, or 14.8 correct responses out of a possible 25. Summary statistics were congruent with the first pilot analysis including: relative real and model alignment among items (Real RMSE = .20, *SD* = 1.32; Model RMSE = .20, *SD* = 1.32) and persons (Real RMSE = .50, *SD* = .65; Model RMSE = .52, *SD* = .67). The items ranged from -2.26 to 2.59 logits, and persons ranged from -6.76 to -.63 logits. Item separation was still high (6.46) with strong item reliability (.98). Person separation was low (1.24) and reliability was acceptable (.60) with a reliability (KR-20) of .62. Person mean performance on the continuum decreased from .11 in the previous analysis including all 35 questions, to -3.69 logits for 25 items. This average person ability indicated that the pilot sample had very low classroom assessment literacy levels as measured by the 25-item CALI. High infit and outfit person values confirmed that low performing persons or difficult items may be related to this decrease. For the items, fit was improved overall from the 35-item CALI, indicated by a lack of point-measure correlation concerns. However, four items had high ZSTD outfit statistics, ranging from 2.3 to 3.6.



```
<more> --------------------- PERSON  -+- ITEM     ----------------- <rare>
  3                                   +                                   3
                                      |T
                                      |  XX
                                      |
  2                                   +  X                                2
                                      |  X
                                      |
                                      |S X
  1                                   +                                   1
                                      |  XXX
                                      |  X
                                      |  XX
  0                                   +M XX                               0
                                      |  X
                                      |  XX
                                   .  |  X
 -1                                   +  XXX                             -1
                                      |S X
                                      |  X
                                      |  X
 -2                                . T+                                  -2
                                  .#  |  XX
                                      |
                           ######### S|T
 -3                         ########  +                                  -3
                         .##########  |
                            .#######  |
                              ###### M|
 -4             .###################  +                                  -4
                             #######  |
                                .### S|
                                ####  |
 -5                               .#  +                                  -5
                                  .# T|
                                   .  |
                                   .  |
 -6                                   +                                  -6
                                      |
                                      |
                                   #  |
 -7                                   +                                  -7
<less> --------------------- PERSON  -+- ITEM     ----------------- <freq>
```

Figure 6. Pilot Sample 25-Item Item-Person Map (*N* = 165). This vertical scale illustrates the placement of persons and items on a continuum representing the latent variable (i.e., assessment knowledge). The left side of the scale presents the logit values, which are a common interval scale created by the Rasch Model with a mean of 0.0. Items are located on the right side of the scale. Items higher on the continuum were more difficult for persons in this sample to endorse, while items lower on the continuum were easier. Persons are placed on the left side of the scale according to their ability and the degree of difficulty of the items

Table 11

Pilot Sample 25-Item CALI Summary of Person Statistics (N = 165)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	14.8	25.0	-3.69	.49	.99/-.1	1.09/.0
P.SD	3.4	.0	.83	.04	.34/1.1	1.12/1.1
S.SD	3.4	.0	.83	.04	.34/1.1	1.12/1.1
Max	24.0	25.0	-.63	.79	4.15/5.8	9.90/4.3
Min	3.0	25.0	-6.76	.47	.42/-2.2	.16/-1.8

Note. Real/Model RMSE = .52/.50; Real/Model True SD = .65/.67; Real/Model Separation = 1.24/1.34; Real/Model Person Reliability = .60/.64; Standard Error of Person Mean = .06; Coefficient Alpha (KR-20) = .62, SEM = 2.09.

Table 12

Pilot Sample 25-Item CALI Summary of Item Statistics (N = 25)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	97.5	165.0	.00	.20	.98/.0	1.11/.6
P.SD	39.3	.0	1.33	.04	.08/.8	.35/1.5
S.SD	40.1	.0	1.36	.04	.09/.8	.36/1.6
Max	153.0	165.0	2.59	.29	1.11/2.0	2.08/3.6
Min	22.0	165.0	-2.26	.17	.75/-1.3	.61/-1.4

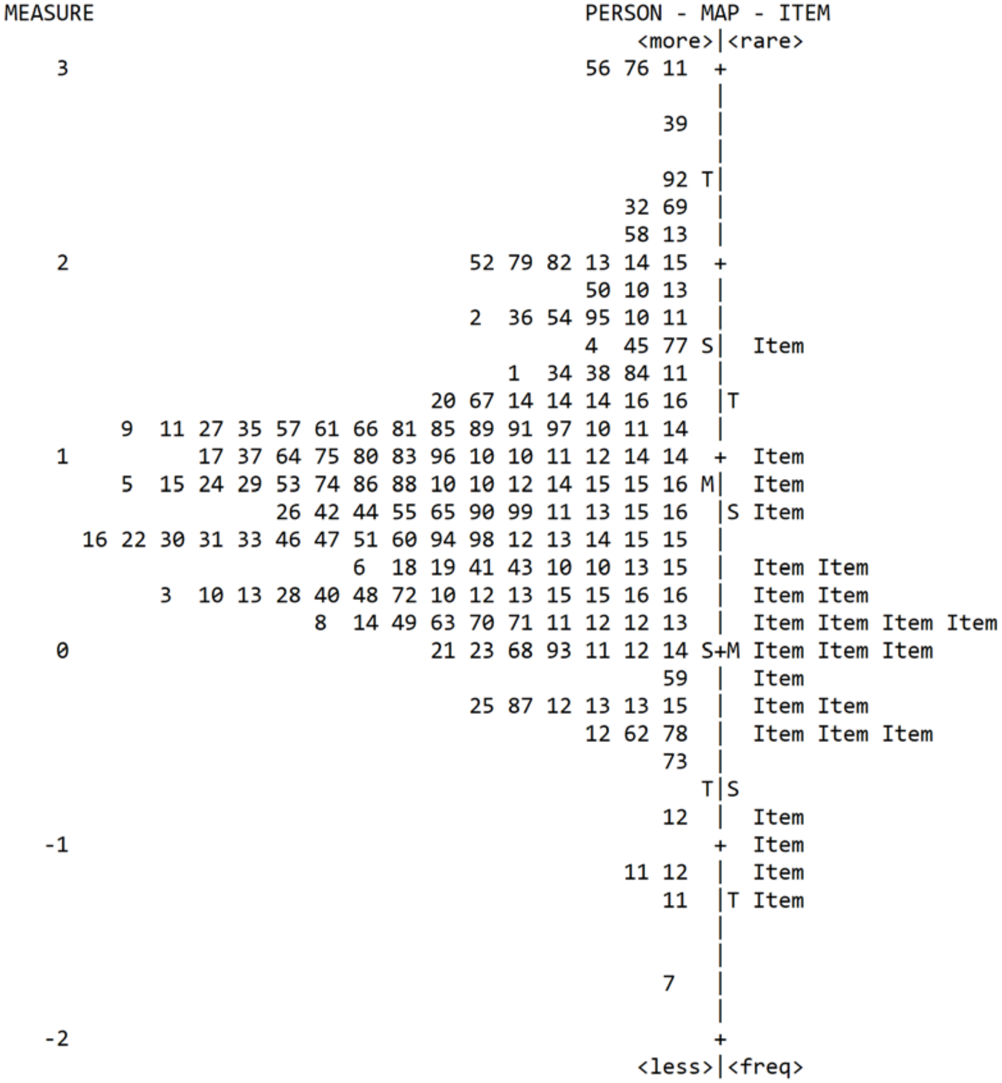
Note. Real/Model RMSE = .20/.20; Real/Model True SD = 1.32/1.32; Real/Model Separation = 6.21/6.52; Real/Model Item Reliability = .90/.98; Standard Error of Item Mean = .27.

Table 13

Pilot Sample 25-Item CALI Misfit Statistics (N = 25)

Item	Measure	Model *SE*	Infit	Outfit	Point-Measure
			MNSQ/ZSTD	MNSQ/ ZSTD	Correlation
22	2.48	.23	1.04/.3	2.08/3.6	.12 (.25)
21	2.59	.24	1.09/.6	2.07/3.4	.09 (.24)
5	1.91	.20	1.05/.5	1.59/3.1	.18 (.28)
15	1.72	.19	.97/-.3	1.37/2.3	.29 (.30)
2	.86	.17	1.11/2.0	1.31/3.4	.24 (.34)
14	.78	.17	1.04/.7	1.17/2.0	.28 (.34)
8	1.27	.17	1.02/.3	1.14/1.3	.32 (.32)
19	.86	.17	1.04/.8	1.12/1.4	.27 (.34)
4	.11	.17	1.09/1.6	1.11/1.4	.30 (.35)
6	-.37	.18	1.01/.1	1.05/.5	.32 (.34)
10	-.40	.18	.98/-.2	1.02/.2	.34 (.34)
18	-.74	.19	.96/.-4	1.01/.1	.35 (.33)
20	.34	.17	1.01/.1	.99/-.1	.34 (.35)
11	.58	.17	.99/-.1	.96/-.5	.36 (.34)
23	-.97	.20	.96/-.3	.99/.0	.34 (.32)
1	-1.82	.25	.97/-.1	.98/.0	.40 (.28)
3	-1.05	.20	.98/-.1	.94/-.3	.33 (.32)
17	-.03	.17	.98/-.3	.96/-.4	.37 (.35)
25	.28	.17	.98/-.3	.96/-.4	.37 (.35)
16	-1.18	.21	.92/-6	.97/-.1	.36 (.31)
9	-1.47	.23	.93/-.4	.85/-.7	.35 (.30)
24	-.40	.18	.91/-1.2	.88/-1.1	.44 (.34)
13	-/93	.20	.86/-1.3	.82/-1.2	.46 (.32)
7	-2.26	.29	.78/-.9	.67/-1.1	.35 (.25)
12	-2.18	.29	.75/-1.1	.61/-1.4	.41 (.26)

Secondly, results from analysis of the 25-item confidence portion of the CALI were investigated. That is, the following results investigated the confidence scores of the pilot sample after removing the ten misfitting items. The total group average score increased to .79 from .64, indicating a slightly above average presence of the latent trait (i.e., assessment confidence). Summary statistics were congruent with the previous reports including: relative real and model alignment among items (Real RMSE = .10, *SD* = .66; Model RMSE = .10, *SD* = .66) and persons (Real RMSE = .29 *SD* = .73; Model RMSE = .26, *SD* = .74). The items ranged from -1.29 to 1.60 logits, and persons ranged from -1.66 to 3.00 logits. Item separation was still high (6.28) with strong item reliability (.98). Person separation was moderate (2.56) and reliability was high (.87) with a Cronbach Alpha reliability of .89. Infit and outfit item statistics were acceptable with the exception of a high maximum infit ZSTD (3.8) and a high maximum outfit ZSTD (4.6) value. The same finding was true for persons with high maximum infit ZSTD (5.1) and outfit ZSTD (5.4) values.



```
MEASURE                                      PERSON - MAP - ITEM
                                                 <more>|<rare>
    3                                        56 76 11  +
                                                       |
                                                   39  |
                                                       |
                                                   92 T|
                                                32 69  |
                                                58 13  |
    2                               52 79 82 13 14 15  +
                                             50 10 13  |
                                    2  36 54 95 10 11  |
                                             4  45 77 S|  Item
                                       1  34 38 84 11  |
                                 20 67 14 14 14 16 16  |T
         9  11 27 35 57 61 66 81 85 89 91 97 10 11 14  |
    1          17 37 64 75 80 83 96 10 10 11 12 14 14  +  Item
         5  15 24 29 53 74 86 88 10 10 12 14 15 15 16 M|  Item
                     26 42 44 55 65 90 99 11 13 15 16  |S Item
      16 22 30 31 33 46 47 51 60 94 98 12 13 14 15 15  |
                           6  18 19 41 43 10 10 13 15  |  Item Item
            3  10 13 28 40 48 72 10 12 13 15 15 16 16  |  Item Item
                        8  14 49 63 70 71 11 12 12 13  |  Item Item Item Item
    0                            21 23 68 93 11 12 14 S+M Item Item Item
                                                   59  |  Item
                                    25 87 12 13 13 15  |  Item Item
                                             12 62 78  |  Item Item Item
                                                   73  |
                                                      T|S
                                                   12  |  Item
   -1                                                  +  Item
                                                11 12  |  Item
                                                   11  |T Item
                                                       |
                                                       |
                                                   7   |
                                                       |
   -2                                                  +
                                                 <less>|<freq>
```

Figure 7. Pilot Sample 25-Item Confidence Item-Person Map (*N* = 165). This vertical scale illustrates the placement of persons and items on a continuum representing the latent variable (i.e., assessment knowledge). The left side of the scale presents the logit values, which are a common interval scale created by the Rasch Model with a mean of 0.0. Items are located on the right side of the scale. Items higher on the continuum were more difficult for persons in this sample to endorse, while items lower on the continuum were easier. Persons are placed on the left side of the scale according to their ability and the degree of difficulty of the items.

Table 14

Pilot Sample 25-Item Confidence Measure Summary of Person Statistics (N = 165)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	66.2	25.0	.79	.26	1.02/-.1	1.01/-.1
P.SD	12.5	.0	.79	.03	.56/1.8	.57/1.8
S.SD	12.5	.0	.79	.03	.56/1.8	.58/1.8
Max	92.0	25.0	3.00	.40	3.03/5.1	3.71/5.4
Min	20.0	25.0	-1.66	.22	.23/-3.8	.23/-4.0

Note. Real/Model RMSE = .29/.26; Real/Model True SD = .73/.74; Real/Model Separation = 2.56/2.88; Real/Model Person Reliability = .87/.89; Standard Error of Person Mean = .06; Coefficient Alpha (KR-20) = .89, SEM = 4.09.

Table 15

Pilot Sample 25-Item Confidence Measure Summary of Item Statistics (N = 25)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	437.1	165.0	.00	.10	1.03/.1	1.02/.1
P.SD	71.4	.0	.67	.01	.23/1.9	.23/2.0
S.SD	72.9	.0	.68	.01	.23/2.0	.24/2.0
Max	555.0	165.0	1.60	.12	1.50/3.8	1.61/4.6
Min	240.0	165.0	-1.29	.09	.77/-2.1	.71/-2.8

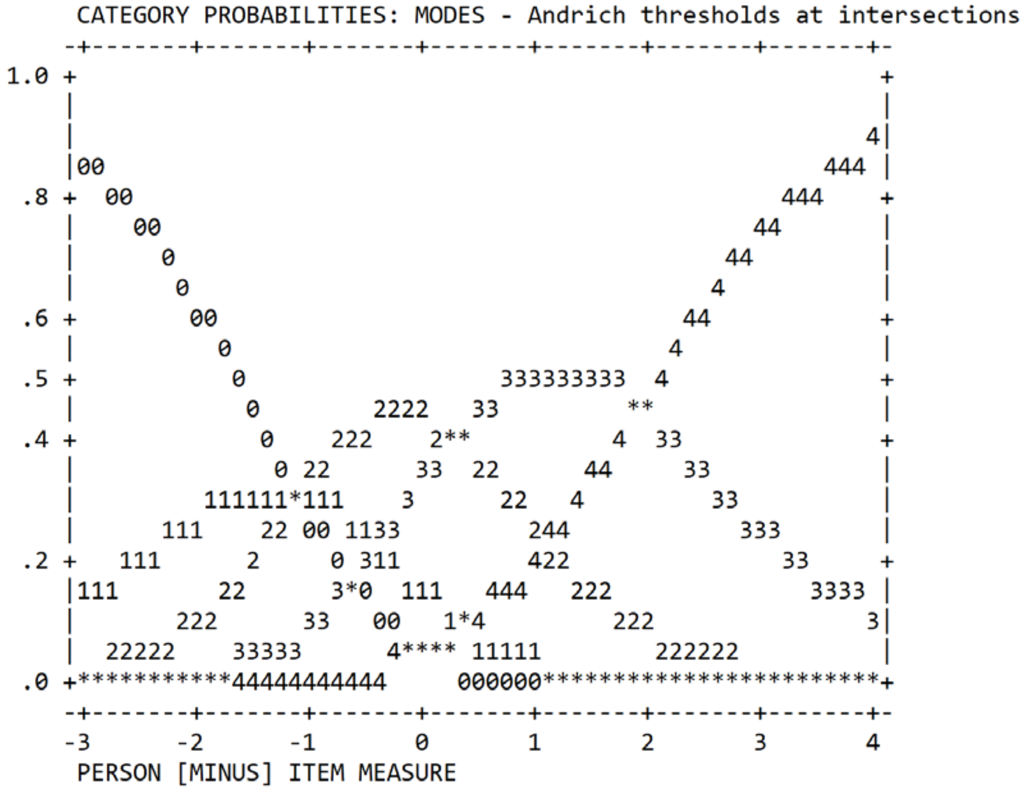
Note. Real/Model RMSE = .10/.10; Real/Model True SD = .66/.66; Real/Model Separation = 6.28/6.64; Real/Model Item Reliability = .98/.98; Standard Error of Item Mean = .14.

Table 16

Pilot Sample 25-Item Confidence Measure Misfit Statistics (N = 25)

Item	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ ZSTD	Point-Measure Correlation
1	-.36	.10	1.50/3.8	1.61/4.6	.37 (.50)
24	.18	.09	1.49/3.8	1.48/3.8	.53 (.53)
23	-.40	.10	1.45/3.5	1.39/3.1.	.47 (.50)
12	-1.29	.12	1.37 (2.8)	1.15 (1.2)	.56 (.44)
22	1.60	.09	1.28 (2.7)	1.34 (3.1)	.32 (.56)
2	.14	.10	1.23/1.9	1.25/2.1	.42 (.53)
13	-.83	.11	1.20/1.7	1.21/1.7	.57 (.47)
17	-.31	.10	1.04/.4	1.08/.7	.58 (.51)
21	.71	.09	1.02/.3	1.05/.5	.55 (.55)
6	-.19	.10	1.02/.2	1.04/.4	.56 (.51)
9	-1.02	.12	1.03/.3	1.00/.0	.46 (.46)
14	.46	.09	.99/-.1	.95/-.5	.48 (.55)
25	.15	.09	.98/-.1	.95/-.4	.52 (.53)
7	-1.20	.12	.97/-.2	.89/-.9	.53 (.44)
20	.20	.09	.96/-.3	.97/-.2	.63 (.53)
10	-.34	.10	.94/-.5	.91/-.7	.58 (.51)
4	.33	.09	.88/-1.1	.88/-1.1	.46 (.54)
8	.36	.09	.81/-1.8	.85/-1.4	.59 (.54)
15	.86	.09	.85/-1.5	.84/-1.5	.52 (.56)
18	.00	.10	.85/-1.3	.85/-1.3	.59 (.52)
19	1.05	.09	.82/-1.9	.84/-1.5	.51 (.56)
5	.30	.09	.78/-2.0	.80/-1.9	.53 (.54)
11	.02	.10	.78/-2.0	.79/-2.0	.62 (.53)
3	-.01	.10	.77/-2.1	.76/-2.3	.53 (.52)
16	-.41	.10	.77/-2.1	.71/-2.8	.62 (.50)

These high values were observed in the item-person map showing 24 persons placed higher on the continuum than the highest item. In other words, it was easy for nearly 20% of the sample to highly endorse all items (i.e., “Completely Confident”), including the most difficult item to endorse. With regards to category structure, the findings were consistent with the 35-item version of the measure. The Andrich Thresholds were monotonically increasing. Additionally, there was limited distinction between the adjacent categories of “Completely Unconfident” and “Somewhat Unconfident” and the adjacent categories of “Somewhat Unconfident” and “Neither Confident Nor Unconfident,” Finally, the majority of responses fell under “Somewhat Confident.”



```
     CATEGORY PROBABILITIES: MODES - Andrich thresholds at intersections
    -+-------+-------+-------+-------+-------+-------+-------+-
1.0 +                                                         +
    |                                                         |
    |                                                        4|
    |00                                                   444 |
 .8 +  00                                              444    +
    |    00                                          44       |
    |      0                                       44         |
    |       0                                     4           |
 .6 +        00                                 44            +
    |          0                               4              |
 .5 +           0                  333333333  4               +
    |            0        2222   33         **                |
 .4 +             0    222    2**          4  33              +
    |              0 22      33  22      44     33            |
    |         111111*111    3      22   4         33          |
    |      111    22 00 1133         244            333       |
 .2 +   111      2     0 311         422               33     +
    |111       22     3*0  111   444   222              3333 |
    |      222      33   00   1*4         222                3|
    |  22222    33333      4**** 11111        222222          |
 .0 +***********44444444444     000000************************+
    -+-------+-------+-------+-------+-------+-------+-------+-
    -3      -2      -1       0       1       2       3       4
     PERSON [MINUS] ITEM MEASURE
```

Figure 8. Pilot Sample 25-Item Confidence Category Probabilities (*N* = 165). This figure indicates the probability of endorsement for each of the Likert scale confidence categories, starting with “Completely Unconfident” on the left and ending with “Complete Confident” on the right. The strength of the peak between each category indicates the discrimination between response levels.

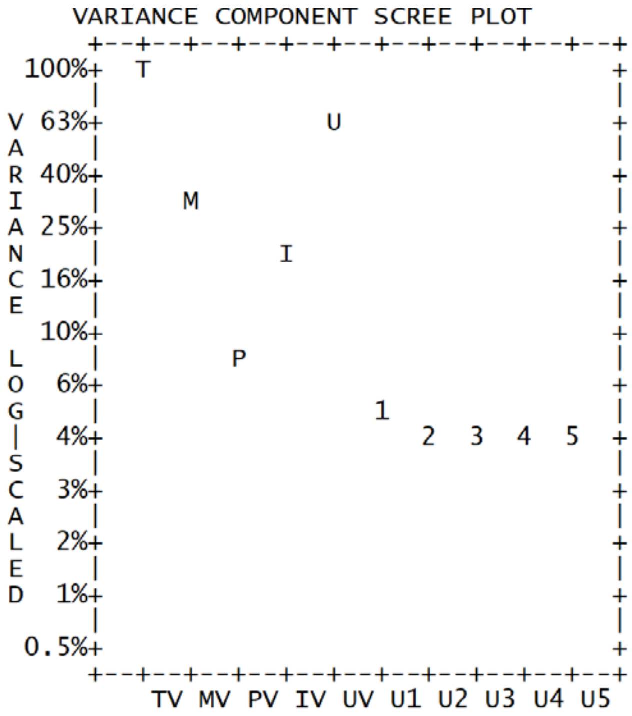
Table 17

Pilot Sample 25-Item Confidence Measure Category Thresholds and Fit Table (N = 25)

Category Label	Observed Count	Observed Average	Sample Expected	Infit MNSQ	Outfit MNSQ	Andrich Threshold	Category Measure
0	180 (4%)	-.64	-.83	1.22	1.29	None	(-2.66)
1	342 (8%)	-.10	-.19	1.12	1.22	-1.14	-1.29
2	1098 (27%)	.29	.38	.91	.89	-1.07	-.21
3	1631 (40%)	.95	.97	1.01	.97	.28	1.21
4	874(21%)	1.76	1.67	.92	.93	1.93	(3.16)

Rasch Principal Components Analysis (PCA)

A Rasch PCA was conducted to examine the dimensionality of the CALI. The total raw variance explained in the observations of these data reported an Eigenvalue 37.041. This equates to 32.5% of the total variance being explained by all observed measures. The raw unexplained total variance was 67.5%. The more items are of equal difficulty and persons are of similar ability (e.g., pre-service teachers at the end of their training), the less variance the measure will explain (Linacre, 2010). All persons accounted for 9.3% of this variance, while items explained 23.2% of the total observed variance. These results are graphically depicted in Figure 9 below.



```
     VARIANCE COMPONENT SCREE PLOT
      +--+--+--+--+--+--+--+--+--+--+--+
  100%+  T                             +
      |                                |
 V 63%+             U                  +
 A    |                                |
 R 40%+                                +
 I    |    M                           |
 A 25%+                                +
 N    |          I                     |
 C 16%+                                +
 E    |                                |
   10%+                                +
 L    |       P                        |
 O  6%+                                +
 G    |                1               |
 |  4%+                   2  3  4  5   +
 S    |                                |
 C  3%+                                +
 A    |                                |
 L  2%+                                +
 E    |                                |
 D  1%+                                +
      |                                |
  0.5%+                                +
      +--+--+--+--+--+--+--+--+--+--+--+
          TV MV PV IV UV U1 U2 U3 U4 U5
```

Figure 9. PCA Figure with Variance and Components. This figure shows the variance accounted for by different elements in the model. Variance is account for by persons (P), items (I), the model (M), uniqueness (U), and the total (T). Additionally, the right aligned numbers indicate the possible components – only component two was strong enough to be considered.

Standardized residuals reported unexplained variance from five contrasts with the first contrast having the strength of 2.2 items, or 5.9% of the observed variance. Additional contrasts included: the second contrast at 1.78 items and 4.8% observed, the third at 1.74 items and 4.7% observed, the fourth contrast at 1.61 items and 4.3% observed, and the fifth contrast at 1.49 items and 4.0% observed. In sum, the variance explained by all items (23.2%,) was only approximately three and a half times the variance explained by the first contrast (5.9%). Notably, the first contrast was the only contrast with an Eigenvalue greater than two. The eigenvalue of the first contrast is 2.2, the smallest amount that could be considered a "dimension" (Linacre, 2010). This indicates a marginally noticeable secondary dimension in the items.

Three item clusters were reported and used to assess differences in items at the top of the difficulty level of the measure versus its opposite (i.e., items possessing less of the latent variable). These clusters also produce two types of correlations, Pearson Correlations and Disattenuated Correlations, where persons are measured for each cluster of items and are correlated with their measures from the other clusters of items (Linacre, 2010). The Pearson Correlation for Clusters 1-3 was very low at .051, Clusters 1-2 approached a moderate level at .432, and Clusters 2-3 was low at .262. This indicates that there was a small to moderate correlation between persons' scores on items in Clusters 1-2. The disattenuated correlation is the observed correlation between two variables when the measurement error has been removed by a statistical operation (Linacre, 2010). Clusters 1-3 have a disattenuated correlation of .119, Clusters 1-2 at 1.00, and Clusters 2-3 at .576. According to Linacre (2010), correlations below .57, as seen in Clusters 1-3 and Clusters 2-3, indicate that person measures on the two item clusters have half as much variance in common as they have independently. In other words, disattenuated correlations of .57 or less indicate there is a different latent variable present.

In relation to Clusters 1-2, any disattenuated correlations greater than .82 generally account for enough of the same variance that the item clusters measure the same latent variable. This analysis provides evidence of a second dimension, likely between items in Clusters 1-3. While, the items Clusters 2-3 are borderline (i.e., have a disattenuated correlation of .576), the Eigenvalue of 1.78 suggests there is only a singular secondary dimension.

Lastly, investigating the items sorted by loading indicated that 12 items positively loaded above zero and the remaining 13 items were negatively loaded on the scale of the latent trait. The positive items ranged from .49 to .02, while the negative loadings ranged from -.02 to -.48. These loadings illustrate a near equal distribution of items across the latent variable. The full summary of loadings and items can be seen below in Table 18.

Table 18

Pilot Sample Rasch Principal Components Analysis (PCA) CALI Item Loadings (N = 25)

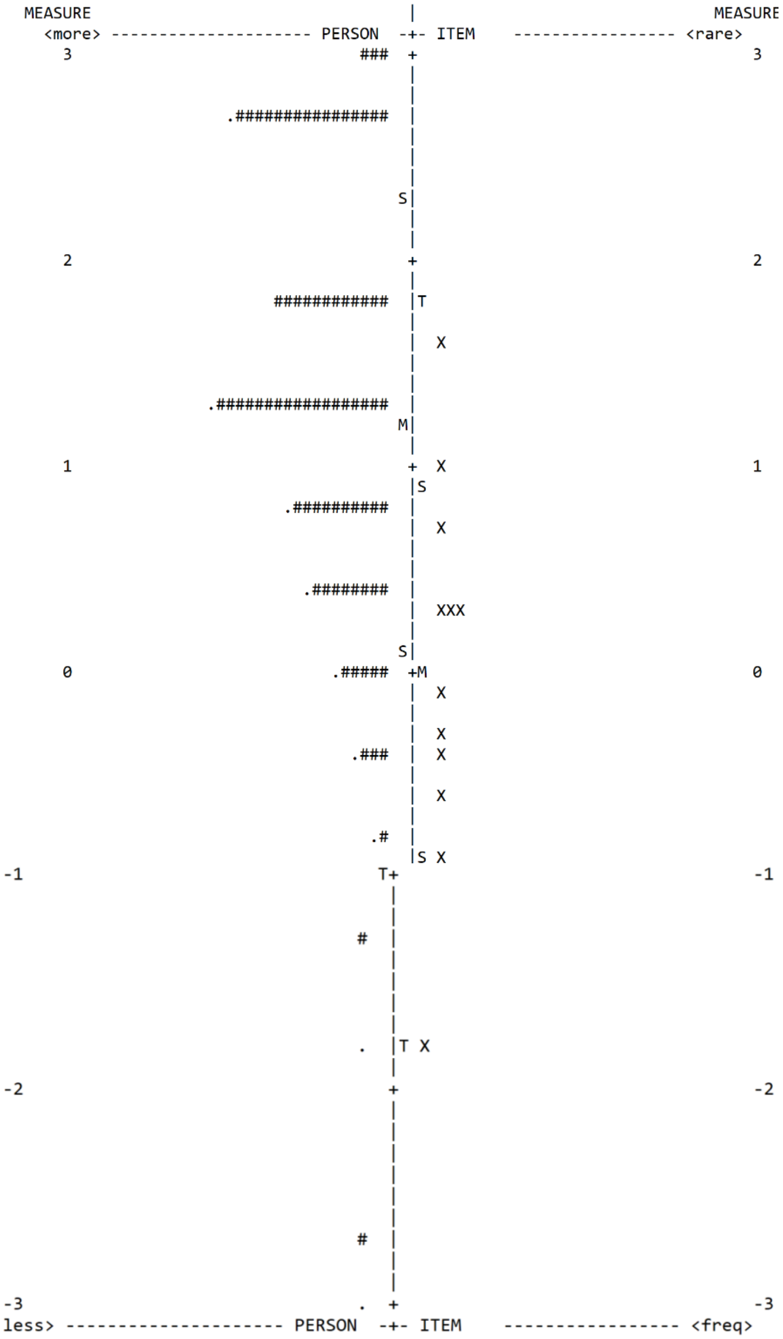
Item*	Content Domain from the *Standards*	Loading
24	Ethical/legal assessment methods and uses of assessment information	.49
13	Using assessment results to inform decisions	.45
23	Ethical/legal assessment methods and uses of assessment information	.41
18	Developing valid grading procedures	.39
10	Administering, scoring and interpreting results	.37
16	Developing valid grading procedures	.37
12	Administering, scoring and interpreting results	.32
6	Developing assessment methods appropriate for instructional decisions	.18
17	Developing valid grading procedures	.14
25	Ethical/legal assessment methods and uses of assessment information	.07
9	Administering, scoring and interpreting results	.05
14	Using assessment results to inform decisions	.02
4	Choosing assessment methods appropriate for instructional decisions	-.02**
2	Choosing assessment methods appropriate for instructional decisions	-.06
1	Choosing assessment methods appropriate for instructional decisions	-.08
8	Developing assessment methods appropriate for instructional decisions	-.09
3	Choosing assessment methods appropriate for instructional decisions	-.21
21	Communicating assessment results to students, parents, other audiences	-.21
15	Developing valid grading procedures	-.22
19	Communicating assessment results to students, parents, other audiences	-.24
22	Ethical/legal assessment methods and uses of assessment information	-.25
5	Choosing assessment methods appropriate for instructional decisions	-.29
20	Communicating assessment results to students, parents, other audiences	-.46

7	Developing assessment methods appropriate for instructional decisions	-.47
11	Administering, scoring and interpreting results	-.48

Note. *See Appendix D for numbering. **Indicates the start of the second dimension.

The above results indicate that the components or content domains of the CALI, as defined by the *Standards,* do not equate to actual, measurable dimensions. Thus, through an investigation of the dimensionality of the CALI, the seven domains in the *Standards* are more broadly categorized into two possible components in the Rasch PCA analysis in this sample. Additionally, and most importantly, although the results suggest some evidence of two components, a unidimensional internal structure appeared equally probable. The items can be seen in Appendix E according to their PCA loading and are investigated further below.

First component Rasch analysis. All 12 multiple-choice items were scored as "Correct" (Coded 1) or "Incorrect" (Coded 0). The initial analysis converged after five iterations. The average score across all participants was 8.6 points or 71.7% correct responses (*SD* = 2.30). Initial investigation of the item-person map revealed the range of items and persons to be between -1.80 and approximately 3.0 logits. Specifically, items fell between -1.80 and 1.60 logits and persons appeared between -.80 and 3.0 logits. Items and persons were largely contained between -.90 and 1.80 logits; however, some items/persons fell outside this range. No items were located above the highest performing person (3.0 logits), indicating that these items were easily endorsed by some participants in this sample. Alternatively, two items were easily answered by all respondents and were located below -1.0 logits or the lowest participant on the scale. One grouping of three items was located at the same difficulty level on the vertical ruler, which may indicate related content.



```
  MEASURE                                  |                               MEASURE
    <more> ---------------------- PERSON  -+- ITEM    ----------------- <rare>
      3                               ###  +                                   3
                                           |
                                           |
                        .################  |
                                           |
                                           |
                                           |
                                          S|
                                           |
                                           |
      2                                    +                                   2
                                           |
                             ############  |T
                                           |
                                           |  X
                                           |
                                           |
                      .##################  |
                                          M|
                                           |
      1                                    +  X                                1
                                           |S
                              .##########  |
                                           |  X
                                           |
                                           |
                                .########  |
                                           |  XXX
                                           |
                                          S|
      0                            .#####  +M                                  0
                                           |  X
                                           |
                                           |  X
                                     .###  |  X
                                           |
                                           |  X
                                           |
                                       .#  |
                                           |S X
-1                                      T+                                     -1
                                         |
                                         |
                                      #  |
                                         |
                                         |
                                         |
                                         |
                                      .  |T X
                                         |
-2                                       +                                     -2
                                         |
                                         |
                                         |
                                         |
                                         |
                                         |
                                      #  |
                                         |
                                         |
-3                                    .  +                                     -3
less> ------------------------- PERSON  -+- ITEM     ------------------ <freq>
```

Figure 10. Pilot Sample 12-Item CALI Component 1 Item-Person Map (*N* = 165). This vertical scale illustrates the placement of persons and items on a continuum representing the latent variable (i.e., assessment knowledge). The left side of the scale presents the logit values, which are a common interval scale created by the Rasch Model with a mean of 0.0. Items are located on the right side of the scale. Items higher on the continuum were more difficult for persons in this sample to endorse, while items lower on the continuum were easier. Persons are placed on the left side of the scale according to their ability and the degree of difficulty of the items.

Item summary statistics revealed relative congruence and minimal misfit data between the real and model values. Therefore, only the real item and person summary statistics are reported. Item RMSE was low (.22, *SD* = .85). Additionally, item separation (3.89) was good and item reliability (.94) was strong, suggesting equal distribution in items across difficulty levels and consistency in item placement. High item separation (i.e., > 3 logits) coupled with high item reliability (i.e., > .9) implies that the person sample is large enough to confirm the item difficulty hierarchy and provide evidence of construct validity (Linacre, 1991). Items spanned the range of -1.81 to 1.61 logits with the expected mean of zero. Reviewing the summarized item infit and outfit statistics indicated a good infit values (MNSQ = 1.11, ZSTD = 1.20) but a high outfit maximum (MNSQ= 1.53, ZSTD = 3.7).

Table 19

Pilot Sample 12-Item CALI Component 1 Summary of Person Statistics (N = 165)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	8.6	12.0	1.27	.83	1.00/.1	.98/-.1
P.SD	2.3	.0	1.27	.27	.20/.7	.58/.7

S.SD	2.3	.0	1.28	.27	.21/.7	.58/.7
Max	12.0	12.0	4.0	1.87	1.62/2.2	6.32/2.7
Min	.0	12.0	-4.03	.62	.62/-1.8	.32/-1.5

Note. Real/Model RMSE = .90/.87; Real/Model True SD = .91/.93; Real/Model Separation = 1.01/1.07; Real/Model Person Reliability = .50/.53; Standard Error of Person Mean = .10; Coefficient Alpha (KR-20) = .64, SEM = 1.40.

Table 20

Pilot Sample 12-Item CALI Component 1 Summary of Item Statistics (N = 25)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	118.2	165.0	.00	.21	.99/.1	.98/.1
P.SD	21.1	.0	.88	.04	.08/.8	.23/1.3
S.SD	22.0	.0	.91	.04	.09/.8	.24/1.4
Max	152.0	165.0	1.61	.33	1.11/1.2	1.53/3.7
Min	73.0	165.0	-1.81	.18	.83/-1.4	.56/-1.8

Note. Real/Model RMSE = .22/.21; Real/Model True SD = .85/.85; Real/Model Separation = 3.89/3.95; Real/Model Item Reliability = .94/.94; Standard Error of Item Mean = .26.

Much like the items, person summary statistics revealed congruence and minimal misfit between the real and model values. The RMSE was .83 (*SD* = .74) with the person mean of 1.20 logits. Thus, the positive mean provided evidence of the presence of classroom assessment literacy within this sample. Person separation (.89) and reliability (.44) were acceptable to low, with the person range of -2.71 to 2.69 logits being slightly restricted. Person infit and outfit statistics also revealed high infit maximum values (MNSQ = 1.62, ZSTD = 2.2) and outfit maximum values (MNSQ = 6.32, ZSTD = 2.70). The person raw score reliability (KR20) was acceptable at .64, indicating moderate reliability in replicating participants' scores across survey administrations (Linacre, 2010).

Individual item misfit was analyzed to further investigate the reported high infit and outfit values. This analysis revealed only one misfitting item with infit or outfit values whose MNSQ exceeded 1.2 (Wright, Linacre, Gustafson, & Martin-Lof, 1994) or ZSTD value exceeding 2 (Wright & Masters, 1982). This was Item 6 (MNS = 1.53, ZSTD = 3.7). All other items had infit and outfit MNSQ values close to 1, and thus, were not misfitting. Lastly, item point-measure correlations did not depart from their expected values and all correlations were positive.

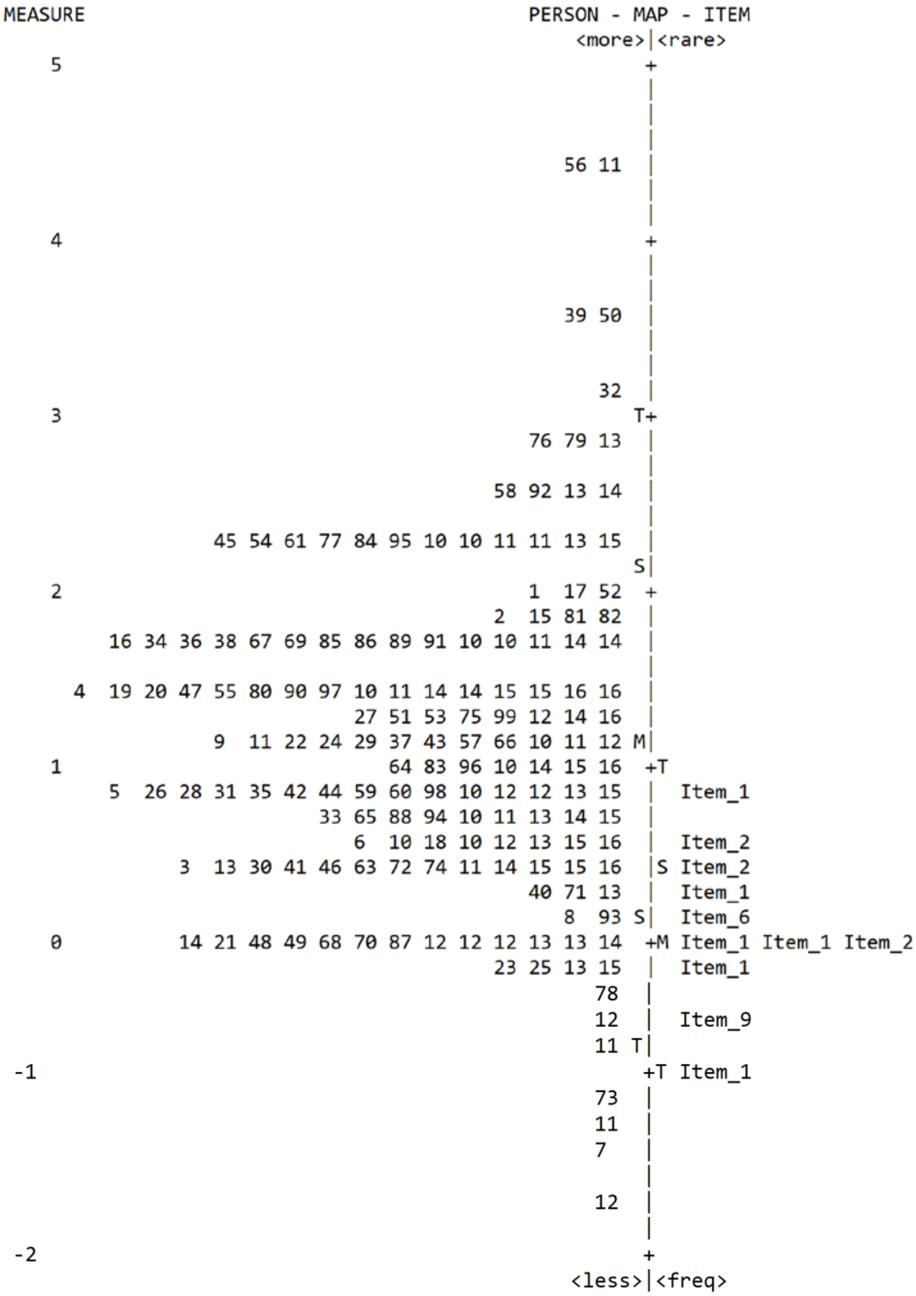
Table 21

Pilot Sample 12-Item CALI Component 1 Item Misfit Statistics (N = 12)

Item	Measure	Model *SE*	Infit	Outfit	Point-Measure
			MNSQ/ZSTD	MNSQ/ ZSTD	Correlation
14	1.61	.18	1.07/1.0	1.53/3.7	.39 (.46)
10	.27	.19	1.05/.6	1.14/1.0	.41 (.45)
6	.31	.19	1.11/1.2	1.13/1.0	.38 (.45)
9	-.95	.25	1.08/.5	.86/-.4	.39 (.41)
25	1.04	.18	1.05/.7	1.03/.4	.43 (.46)
18	-.11	.20	.96/-.4	1.04/.3	.46 (.44)
17	.68	.18	1.02/.3	1.02/.2	.45 (.46)
23	-.37	.21	1.01/.1	.87/-.6	.45 (.43)
16	-.61	.23	.95/-.3	.96/-.1	.45 (.42)
13	-.33	.21	.91/-.8	.79/-1.1	.50 (.44)
24	.27	.19	.88/-1.4	.76/-1.8	.55 (.45)
12	-1.81	.33	.83/-.6	.56/-1.0	.48 (.38)

First component confidence Rasch analysis. The Rasch Rating Scale Model (RSM) was used to analyze these data as all 12 Component 1 confidence items required a response on a 5-point Likert-scale, ranging from "Complete Unconfident" (Coded 0) to

“Completely Confident” (Coded 4). The analysis converged after fourteen iterations. Initial investigation of the item-person map (Figure 11) revealed a large cluster of items between -1.0 and approximately 0.5 logits. The positioning for participants was between -0.5 and 2.0 logits.



```
MEASURE                                      PERSON - MAP - ITEM
                                                 <more>|<rare>
    5                                                  +
                                                       |
                                                       |
                                                       |
                                                56 11  |
                                                       |
                                                       |
    4                                                  +
                                                       |
                                                       |
                                                39 50  |
                                                       |
                                                       |
                                                   32  |
    3                                                 T+
                                             76 79 13  |
                                                       |
                                          58 92 13 14  |
                                                       |
                  45 54 61 77 84 95 10 10 11 11 13 15  |
                                                      S|
    2                                        1  17 52  +
                                          2  15 81 82  |
         16 34 36 38 67 69 85 86 89 91 10 10 11 14 14  |
                                                       |
      4  19 20 47 55 80 90 97 10 11 14 14 15 15 16 16  |
                              27 51 53 75 99 12 14 16  |
                  9  11 22 24 29 37 43 57 66 10 11 12 M|
    1                           64 83 96 10 14 15 16  +T
         5  26 28 31 35 42 44 59 60 98 10 12 12 13 15  |  Item_1
                           33 65 88 94 10 11 13 14 15  |
                              6  10 18 10 12 13 15 16  |  Item_2
               3  13 30 41 46 63 72 74 11 14 15 15 16  |S Item_2
                                             40 71 13  |  Item_1
                                                8  93 S|  Item_6
    0          14 21 48 49 68 70 87 12 12 12 13 13 14  +M Item_1 Item_1 Item_2
                                          23 25 13 15  |  Item_1
                                                   78  |
                                                   12  |  Item_9
                                                   11 T|
 -1                                                    +T Item_1
                                                   73  |
                                                   11  |
                                                   7   |
                                                       |
                                                   12  |
                                                       |
 -2                                                    +
                                                 <less>|<freq>
```

Figure 11. Pilot Sample 12-Item CALI Component 1 Confidence Item-Person Map (*N* = 165). This vertical scale illustrates the placement of persons and items on a continuum representing the latent variable (i.e., assessment confidence). The left side of the scale presents the logit values, which are a common interval scale created by the Rasch Model with a mean of 0.0. Items are located on the right side of the scale. Items higher on the continuum were more difficult for persons in this sample to endorse, while items lower on the continuum were easier. Persons are placed on the left side of the scale according to their ability and the degree of difficulty of the items.

Item summary statistics revealed relative congruence and minimal misfit data between the real and model values. Therefore, only the real item and person summary statistics are reported. RMSE was low (.11, *SD* = .48). Additionally, item separation (4.47) was high and reliability (.84) was strong, suggesting equal distribution in items across difficulty levels and consistency in item placement. High item separation (i.e., > 3 logits) coupled with high item reliability (i.e., > .9) implies that the person sample is large enough to confirm the item difficulty hierarchy and provide evidence of construct validity (Linacre, 1991). Items spanned the range of -.97 to .80 logits with the expected mean of zero. Reviewing the summarized item infit and outfit statistics indicated a high infit maximum value (MNSQ = 1.38, ZSTD = 3.0) and outfit maximum (MNSQ = 1.44, ZSTD = 3.4).

Much like items, person summary statistics revealed congruence and minimal misfit between the real and model values. The RMSE was .46 (*SD* = .86) with higher person mean (1.10) indicating the cluster observed in the item-person map. Thus, the positive mean provided evidence of the presence of assessment confidence within this sample. Person separation (1.89) and reliability (.78) were acceptable; with a person range of -1.72 to 4.39 logits. Person infit and outfit statistics also revealed high infit maximum (MNSQ = 4.35, ZSTD = 4.4) and outfit maximum (MNSQ = 4.67, ZSTD = 4.8). The person raw score reliability (Coefficient Alpha) was good at .84, indicating reliability in replicating participants' scores across survey administrations (Linacre, 2010).

Table 22

Pilot Sample 12-Item Component 1 Confidence Measure Person Statistic (N = 165)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	34.4	12.0	1.10	.40	1.01/-.1	1.01/-.1
P.SD	6.8	.0	.98	.10	.67/1.5	.67/1.5
S.SD	6.8	.0	.98	.10	.67/1.5	.67/1.5
Max	47.0	12.0	4.39	1.03	4.35/4.4	4.67/4.8
Min	8.0	12.0	-1.72	.29	.12/-3.3	.12/-3.3

Note. Real/Model RMSE = .46/.41; Real/Model True SD = .86/.89; Real/Model Separation = 1.89/2.15; Real/Model Person Reliability = .78/.82; Standard Error of Person Mean = .08; Coefficient Alpha (KR-20) = .84, SEM = 2.76.

Table 23

Pilot Sample 12-Item Component 1 Confidence Measure Item Statistic (N = 12)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	473.6	165.0	.00	.10	1.02/.1	1.01/.0
P.SD	46.1	.0	.50	.01	.19/1.6	.20/1.7
S.SD	48.1	.0	.52	.01	.20/1.7	.21/1.7
Max	555.0	165.0	.80	.12	1.38/3.0	1.44/3.4
Min	391.0	165.0	-.97	.09	.70/-2.8	.65/-3.3

Note. Real/Model RMSE = .11/.10; Real/Model True SD = .48/.49; Real/Model Separation = 4.47/4.67; Real/Model Item Reliability = .95/.96; Standard Error of Item Mean = .15.

Individual item fit was analyzed to further investigate the summary reported infit and outfit MNSQ values. This analysis revealed Item 24 had a slightly elevated MNSQ infit or outfit values greater than 1.30. This is the suggested value for fit by Smith, Schumacker, and Bush (1995) for a sample of less than 500. Additionally, the MNSQ values for these items did not exceed 1.86 which is also under the 2/-2 value suggested by Smith (1992). All other items had infit and outfit MNSQ values close to 1. Lastly, item point-measure correlations did not depart from their expected values and all correlations were positive.

Table 24

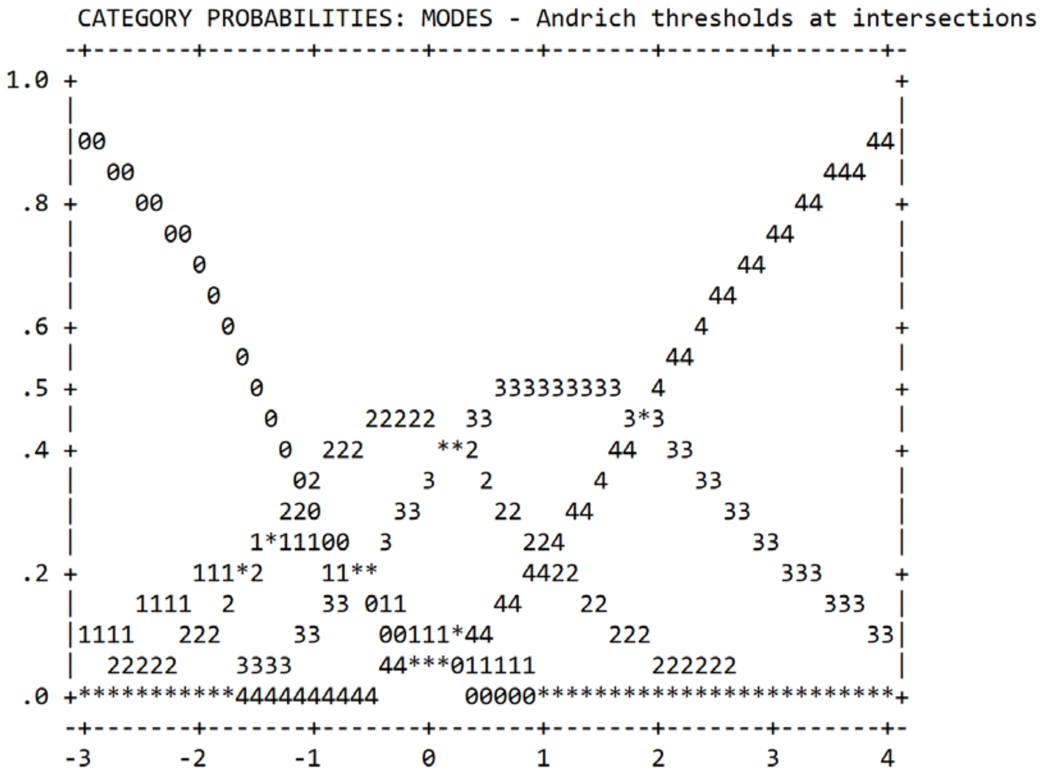
Pilot Sample 12-Item Component 1 Confidence Measure Misfit Statistics (N = 12)

Item	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ ZSTD	Point-Measure Correlation
24	.52	.09	1.38/3.0	1.44/3.4	.56 (.60)
23	-.07	.10	1.28/2.2	1.29/2.3	.53 (.57)
12	-.97	.12	1.20/1.6	1.00/.0	.60 (.50)
9	-.69	.12	1.08/.7	1.06/.5	.47 (.52)
13	-.50	.11	1.06/.6	1.01/.1	.62 (.54)
6	.15	.10	1.03/.3	1.04/.4	.56 (.58)
14	.80	.09	.97/-.3	1.01/.1	.51 (.58)
10	-.01	.10	.96/-.3	.99/.0	.56 (.57)
25	.49	.09	.95/-.4	.94/-.5	.56 (.60)
17	.03	.10	.93/-.5	.92/-.6	.64 (.58)
18	.33	.10	.74/-2.4	.76/-2.2	.64 (.59)
16	-.08	.10	.70/-2.8	.65/-3.3	.65 (.57)

Next, the summary of category structure and observed average of endorsements was investigated. The observed frequencies and percentages for each possible response category revealed higher endorsement for “Neither Confident Nor Unconfident” at 23%

of all responses and “Mostly Unconfident” at 4% of responses. The “Completely Unconfident” category only accounted for 3% of responses, while “Mostly Confident” had 41% and “Completely Confident” had 29%. This pattern of endorsements supports the positive cluster of persons above the mean on the item-person map, indicating slightly above average assessment confidence.

Infit and outfit MNSQ values appeared normal and did not markedly exceed 1.30. Also, the Andrich thresholds (i.e., difficulty between endorsing each response option) displayed a monotonically increasing average. The greatest increase was between “Mostly Confident” (.1.13) and “Completely Confident” (3.08). The large logit gap between these categories demonstrated the relative ease in selecting between the “Mostly Confident” and “Completely Confident” categories compared to other adjacent categories. As shown in Figure 12 below, the lower three Likert categories have closer and less defined crests between “Completely Unconfident,” “Mostly Unconfident,” and “Neither Confident nor Unconfident.”



```
     CATEGORY PROBABILITIES: MODES - Andrich thresholds at intersections
    -+-------+-------+-------+-------+-------+-------+-------+-
1.0 +                                                         +
    |                                                         |
    |00                                                     44|
    |  00                                                444  |
 .8 +    00                                            44     +
    |      00                                        44       |
    |        0                                     44         |
    |         0                                  44           |
 .6 +          0                                4             +
    |           0                             44              |
 .5 +            0                333333333  4                +
    |             0      22222  33         3*3                |
 .4 +              0  222     **2         44  33              +
    |               02       3   2       4      33            |
    |              220     33     22   44         33          |
    |            1*11100  3         224             33        |
 .2 +        111*2    11**          4422              333     +
    |    1111  2      33 011      44    22               333  |
    |1111   222     33    00111*44        222               33|
    |  22222    3333      44***011111        222222           |
 .0 +***********4444444444      00000*************************+
    -+-------+-------+-------+-------+-------+-------+-------+-
    -3      -2      -1       0       1       2       3       4
```

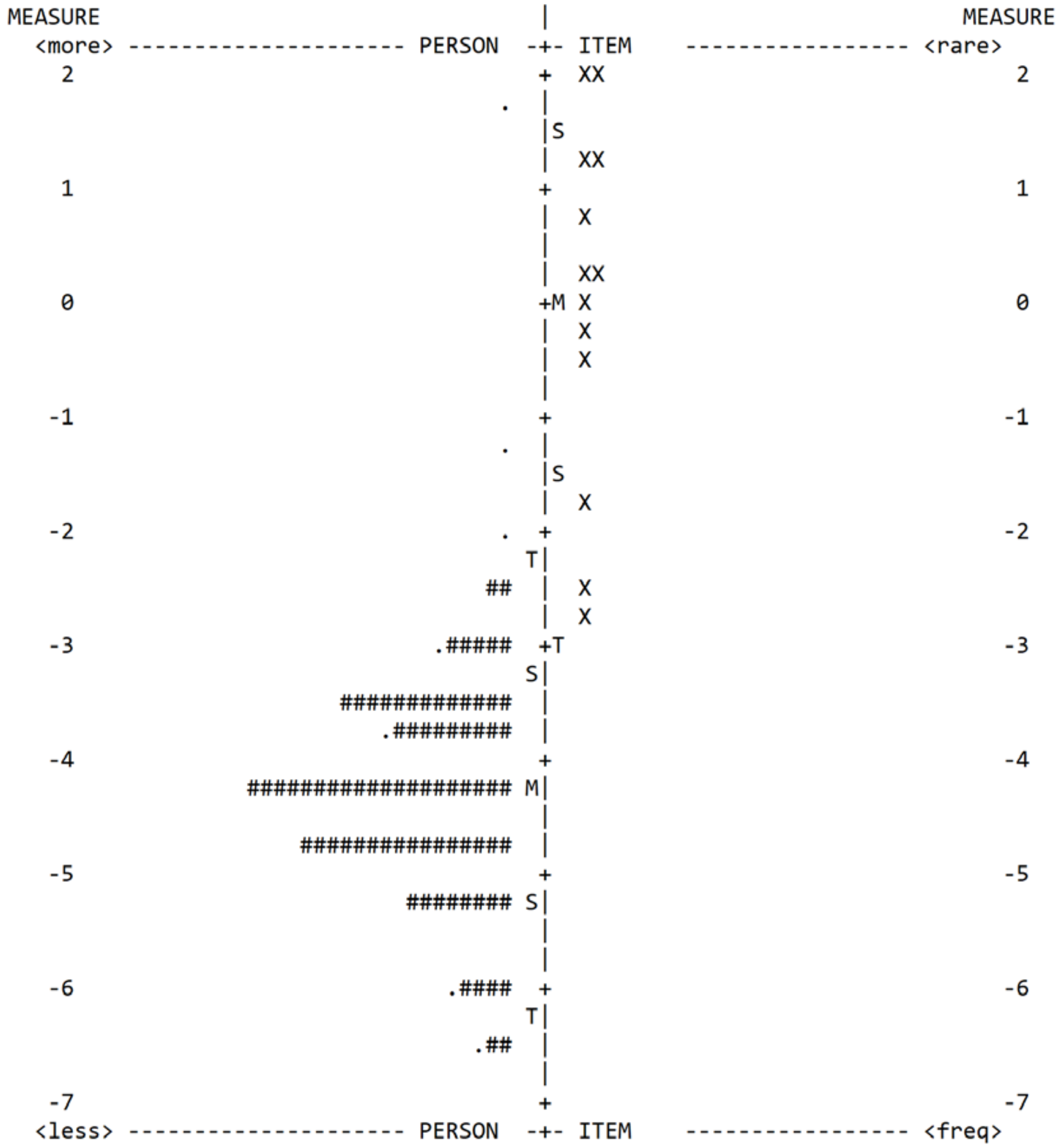
Figure 12. Pilot Sample 12-Item Component 1 Confidence Category Probabilities (*N* = 165). This figure indicates the probability of endorsement for each of the Likert scale confidence categories, starting with “Completely Unconfident” on the left and ending with “Complete Confident” on the right. The strength of the peak between each category indicates the discrimination between response levels.

Table 25

Pilot Sample 12-Item Component 1 Confidence Measure Thresholds and Fit Table (N = 12)

Category Label	Observed Count	Observed Average	Sample Expected	Infit MNSQ	Outfit MNSQ	Andrich Threshold	Category Measure
0	68 (3%)	-.64	-.76	1.15	1.25	None	(-2.43)
1	88 (4%)	.06	-.09	1.18	1.24	-.67	-1.24
2	446 (23%)	.45	.45	.89	.89	-1.40	-.27
3	809 (41%)	1.11	1.11	1.06	.97	.22	1.13
4	569(29%)	1.98	1.98	.97	.97	1.86	(3.08)

Second component Rasch analysis. The Rasch Dichotomous Model was used to analyze these data. All 13 multiple-choice items were scored as “Correct” (Coded 1) or “Incorrect” (Coded 0). The initial analysis converged after five iterations. The average score across all participants was 6.2 points or 51.67% correct responses (*SD* = 2.00). Initial investigation of the item-person map revealed the range of items and persons to be between -6.50 and approximately 2.0 logits. Specifically, Items fell between -2.80 and 2.00 logits and Persons appeared between -6.00 and -2.50 logits. Items and persons had not corresponding clustering. Most items were located above the highest performing person (-2.50 logits), indicating that these items were too difficult to endorse by most participants in this sample. No more than two items were located at the same difficulty level on the vertical ruler, which indicated a variety of content and difficulty.



```
MEASURE                                 |                               MEASURE
  <more> --------------------- PERSON  -+- ITEM    ----------------- <rare>
    2                                   +  XX                               2
                                     .  |
                                        |S
                                        |  XX
    1                                   +                                   1
                                        |  X
                                        |
                                        |  XX
    0                                   +M X                                0
                                        |  X
                                        |  X
                                        |
   -1                                   +                                  -1
                                     .  |
                                        |S
                                        |  X
   -2                                .  +                                  -2
                                       T|
                                    ##  |  X
                                        |  X
   -3                           .#####  +T                                 -3
                                       S|
                         #############  |
                            .#########  |
   -4                                   +                                  -4
                  #################### M|
                                        |
                      ################  |
   -5                                   +                                  -5
                              ######## S|
                                        |
                                        |
   -6                            .####  +                                  -6
                                       T|
                                   .##  |
                                        |
   -7                                   +                                  -7
  <less> --------------------- PERSON  -+- ITEM    ----------------- <freq>
```

Figure 13. Pilot Sample 13-Item Component *2* Item-Person Map (*N* = 165). This vertical scale illustrates the placement of persons and items on a continuum representing the latent variable (i.e., assessment confidence). The left side of the scale presents the logit values, which are a common interval scale created by the Rasch Model with a mean of 0.0. Items are located on the right side of the scale. Items higher on the continuum were more difficult for persons in this sample to endorse, while items lower on the continuum were easier. Persons are placed on the left side of the scale according to their ability and the degree of difficulty of the items.

Item summary statistics demonstrated relative congruence and minimal misfit data between the real and model values. Therefore, only the real item and person summary statistics are reported. RMSE was low (.21, *SD* = 1.49). Additionally, item separation (7.12) was high and item reliability (.98) was strong, suggesting equal distribution in items across difficulty levels and consistency in item placement. High item separation (i.e., > 3 logits) coupled with high item reliability (i.e., > .9) implies that the person sample is large enough to confirm the item difficulty hierarchy and provide evidence of construct validity (Linacre, 1991). Items spanned the range of -2.87 to 2.09 logits with the expected mean of zero. Reviewing the summarized item infit and outfit statistics indicated a good infit values (MNSQ = 1.20, ZSTD = 1.20) but an elevated outfit maximum (MNSQ= 1.88, ZSTD = 2.80).

Table 26

Pilot Sample 13-Item CALI Component 2 Summary of Person Statistics (N = 165)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	6.2	13.0	-4.25	.70	1.01/.0	1.09/.0
P.SD	2.0	.0	1.04	.06	.46/1.0	1.14/.9
S.SD	2.0	.0	1.05	.06	.46/1.1	1.14/1.0
Max	15.0	13.0	1.81	.93	4.04/3.5	9.87/3.2
Min	2.0	13.0	-6.60	.66	.45/-2.0	.21/-1.3

Note. Real/Model RMSE = .76/.70; Real/Model True SD = .71/.77; Real/Model Separation = .93/1.10; Real/Model Person Reliability = .46/.55; Standard Error of Person Mean = .08; Coefficient Alpha (KR-20) = .44, SEM = 1.51.

Table 27

Pilot Sample 13-Item CALI Component 2 Summary of Item Statistics (N = 13)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	78.4	165.0	.00	.20	.98/.0	1.09/.4
P.SD	42.5	.0	1.51	.04	.10/.7	.31/1.2
S.SD	44.2	.0	1.57	.04	.10/.7	.32/1.2
Max	153.0	165.0	2.09	.28	1.20.1.2	1.88/2.8
Min	22.0	165.0	-2.87	.17	.81/-1.3	.66/-1.2

Note. Real/Model RMSE = .21/.21; Real/Model True SD = 1.49/1.49; Real/Model Separation = 7.12/7.23; Real/Model Item Reliability = .98/.98; Standard Error of Item Mean = .44.

Similar to the items, person summary statistics revealed congruence and minimal misfit between the real and model values. The RMSE was .76 (*SD* = .71) with the person mean of 6.20 or -4.25 logits. The negative logit position of the mean provided evidence of the absence of classroom assessment literacy within this sample on these items. Person separation (0.93) and reliability (.46) were acceptable to low, with the person range of -6.60 to 1.81 logits being slightly restricted. Person infit and outfit statistics also revealed high infit maximum values (MNSQ = 4.04, ZSTD = 3.5) and outfit maximum values (MNSQ = 9.87, ZSTD = 3.20). The person raw score reliability (KR-20) was low to moderate at .44, indicating low to moderate reliability in replicating participants' scores across survey administrations (Linacre, 2010).

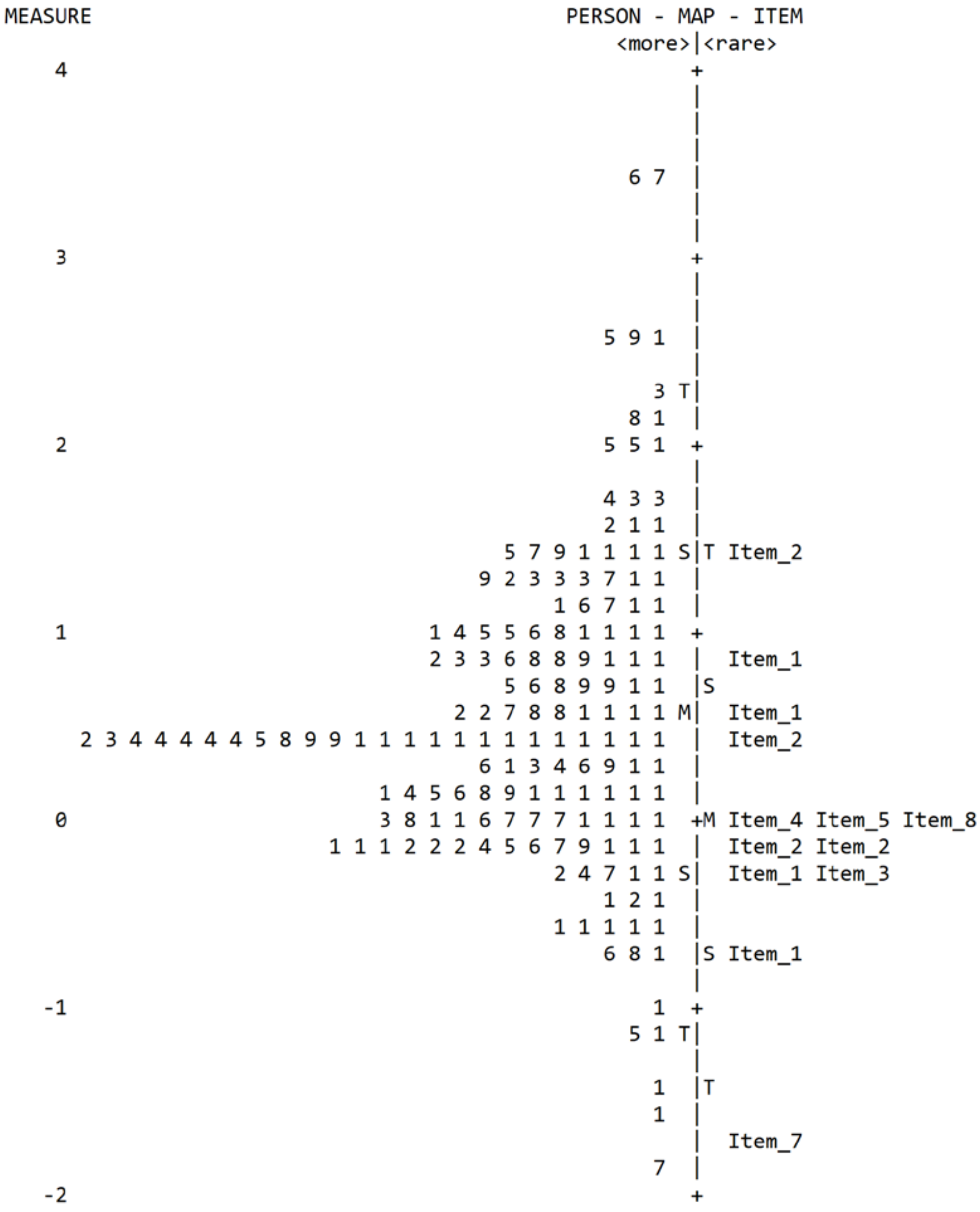
Individual item misfit was analyzed to further investigate the reported high infit and outfit values. This analysis revealed several misfitting items with infit or outfit values whose MNSQ exceeded 1.2 (Wright, Linacre, Gustafson, & Martin-Lof, 1994) or ZSTD value exceeding 2 (Wright & Masters, 1982). All ill-fitting items reported accepted infit statistics but high outfit values. These items included: Item 13 (MNSQ = 1.88, ZSTD = 2.8), Item 12 (MNSQ = 1.43, ZSTD = 1.5), and Item 5 (MNSQ = 1.42, ZSTD = 2.1). All other items had infit and outfit MNSQ values close to 1, and thus, were not misfitting. Lastly, item point-measure correlations did not depart from their expected values and all correlations were positive.

Table 28

Pilot Sample 13-Item CALI Component 2 Item Misfit Statistics (N = 13)

Item	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ ZSTD	Point-Measure Correlation
22	1.97	.24	1.03/.2	1.88/2.8	.23 (.34)
21	2.09	.24	1.20/1.2	1.43/1.5	.15 (.33)
5	1.36	.20	1.04/.4	1.42/2.1	.28 (.36)
8	.69	.18	1.03/.4	1.10/.8	.43 (.38)
2	.27	.17	1.03/.5	1.07/.8	.43 (.38)
19	.27	.17	1.01/.2	1.07/.8	.36 (.38)
15	1.17	.19	.95/-.5	1.05/.4	.37 (.36)
11	-.02	.17	1.03/.6	.98/-.2	.36 (.38)
20	-.28	.17	.99/-.1	1.01/.2	.37 (.38)
4	-.51	.17	.98/-.2	.96/-.4	.47 (.38)
3	-1.69	.20	.85/-1.3	.84/-.9	.42 (.34)
1	-2.45	.25	.84/-.9	.75/-1.0	.47 (.32)
7	-2.87	.28	.81/-.9	.66/-1.2	.28 (.31)

Second component confidence Rasch analysis. The Rasch Rating Scale Model (RSM) was used to analyze these data as all 13 Component 2 confidence items required a response on a 5-point Likert-scale, ranging from “Complete Unconfident” (Coded 0) to “Completely Confident” (Coded 4). The analysis converged after sixteen iterations. Initial investigation of the item-person map (Figure 14) revealed a large cluster of items between -0.5 and approximately 0.5 logits. The positioning for participants was between -0.5 and 1.5 logits.



```
MEASURE                                      PERSON - MAP - ITEM
                                                  <more>|<rare>
    4                                                   +
                                                        |
                                                        |
                                                        |
                                                   6 7  |
                                                        |
                                                        |
    3                                                   +
                                                        |
                                                        |
                                                 5 9 1  |
                                                        |
                                                     3 T|
                                                   8 1  |
    2                                            5 5 1  +
                                                        |
                                                 4 3 3  |
                                                 2 1 1  |
                                         5 7 9 1 1 1 1 S|T Item_2
                                       9 2 3 3 3 7 1 1  |
                                             1 6 7 1 1  |
    1                              1 4 5 5 6 8 1 1 1 1  +
                                   2 3 3 6 8 8 9 1 1 1  |  Item_1
                                         5 6 8 9 9 1 1  |S
                                     2 2 7 8 8 1 1 1 1 M|  Item_1
       2 3 4 4 4 4 4 5 8 9 9 1 1 1 1 1 1 1 1 1 1 1 1 1  |  Item_2
                                       6 1 3 4 6 9 1 1  |
                               1 4 5 6 8 9 1 1 1 1 1 1  |
    0                          3 8 1 1 6 7 7 7 1 1 1 1  +M Item_4 Item_5 Item_8
                           1 1 1 2 2 2 4 5 6 7 9 1 1 1  |  Item_2 Item_2
                                             2 4 7 1 1 S|  Item_1 Item_3
                                                 1 2 1  |
                                             1 1 1 1 1  |
                                                 6 8 1  |S Item_1
                                                        |
   -1                                                1  +
                                                   5 1 T|
                                                        |
                                                     1  |T
                                                     1  |
                                                        |  Item_7
                                                     7  |
   -2                                                   +
```

Figure 14. Pilot Sample 13-Item Component 2 Confidence Item-Person Map (*N* = 165). This vertical scale illustrates the placement of persons and items on a continuum representing the latent variable (i.e., assessment confidence). The left side of the scale presents the logit values, which are a common interval scale created by the Rasch Model

with a mean of 0.0. Items are located on the right side of the scale. Items higher on the continuum were more difficult for persons in this sample to endorse, while items lower on the continuum were easier. Persons are placed on the left side of the scale according to their ability and the degree of difficulty of the items.

Item summary statistics revealed relative congruence and minimal misfit data between the real and model values. Therefore, only the real item and person summary statistics are reported. RMSE was low (.11, *SD* = .72). Additionally, item separation (6.84) was high and item reliability (.98) was strong, suggesting equal distribution in items across difficulty levels and consistency in item placement. High item separation (i.e., > 3 logits) coupled with high item reliability (i.e., > .9) implies that the person sample is large enough to confirm the item difficulty hierarchy and provide evidence of construct validity (Linacre, 1991). Items spanned the range of -1.67 to 1.44 logits with the expected mean of zero. Reviewing the summarized item infit and outfit statistics indicated a high infit maximum value (MNSQ = 1.55, ZSTD = 4.1) and outfit maximum (MNSQ = 1.78, ZSTD = 5.7).

Much like items, person summary statistics revealed congruence and minimal misfit between the real and model values. The RMSE was .40 (*SD* = .74) with higher person mean (.54) indicating the cluster observed in the item-person map. Thus, the positive mean provided evidence of the presence of assessment confidence within this sample. Person separation (1.83) and reliability (.77) were acceptable; with a person range of -1.89 to 3.36 logits. Person infit and outfit statistics also revealed high infit maximum (MNSQ = 3.82, ZSTD = 4.6) and outfit maximum (MNSQ = 5.17, ZSTD = 4.5). The person raw score reliability (Coefficient Alpha) was good at .80, indicating reliability in replicating participants' scores across survey administrations (Linacre, 2010).

Table 29

Pilot Sample 13-Item Component 2 Confidence Measure Summary of Person Statistics (N = 165)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	31.8	13.0	.54	.36	1.02/-.1	1.03/-.1
P.SD	6.6	.0	.84	.04	.64/1.5	.68/1.4
S.SD	6.6	.0	.85	.04	.64/1.5	.69/1.4
Max	48.0	13.0	3.36	.58	3.82/4.6	5.17/4.5
Min	10.0	13.0	-1.89	.32	.24/-2.7	.23/-2.7

Note. Real/Model RMSE = .40/.36; Real/Model True SD = .74/.76; Real/Model Separation = 1.83/2.12; Real/Model Person Reliability = .77/.82; Standard Error of Person Mean = .07; Coefficient Alpha (KR-20) = .80, SEM = 2.96.

Table 30

Pilot Sample 13-Item Component 2 Confidence Measure Summary of Item Statistics (N = 13)

Statistic	Total Score	Count	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ZSTD
M	403.4	165.0	.00	.10	1.02/.0	1.03/.2
P.SD	74.1	.0	.73	.01	.23/2.0	.28/2.3
S.SD	77.1	.0	.76	.01	.24/2.1	.29/2.4
Max	549.0	165.0	1.44	.13	1.55/4.1	1.78/5.7
Min	240.0	165.0	-1.67	.09	.76/-2.3	.75/-2.4

Note. Real/Model RMSE = .11/.10; Real/Model True SD = .72/.72; Real/Model Separation = 6.84/7.23; Real/Model Item Reliability = .98/.98; Standard Error of Item Mean = .21.

Individual item fit was analyzed to further investigate the summary reported infit and outfit MNSQ values. This analysis revealed Item 1 and Item 22 had a slightly elevated MNSQ infit or outfit values greater than 1.30. This is the suggested value for fit by Smith, Schumacker, and Bush (1995) for a sample of less than 500. Additionally, the MNSQ values for these items did not exceed 1.86 which is also under the 2/-2 value suggested by Smith (1992). All other items had infit and outfit MNSQ values close to 1. Lastly, item point-measure correlations did not depart from their expected values and all correlations were positive.

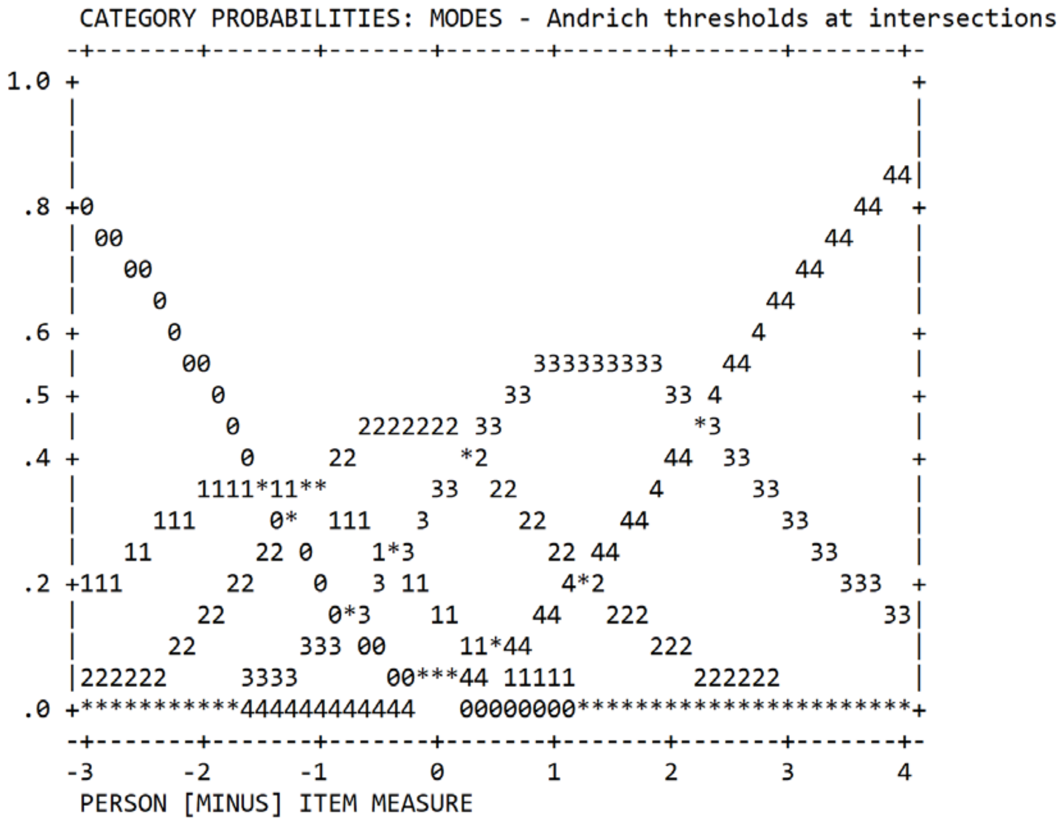
Table 31

Pilot Sample 13-Item Component 2 Confidence Measure Misfit Statistics (N = 13)

Item	Measure	Model *SE*	Infit MNSQ/ZSTD	Outfit MNSQ/ ZSTD	Point-Measure Correlation
1	-.75	.11	1.55/4.1	1.78/5.7	.42 (.50)
22	1.44	.09	1.32/3.0	1.41/3.6	.36 (.57)
2	-.18	.10	1.19/1.6	1.18/1.6	.51 (.53)
21	.44	.09	1.11/1.1	1.12/1.1	.56 (.55)
20	-.12	.10	1.11/1.1	1.11/1.0	.60 (.53)
7	-1.67	.13	1.10/.9	.99/.0	.49 (.44)
8	.06	.10	.88/-1.0	.92/-.7	.60 (.54)
11	-.33	.10	.09/-.8	.89/-.9	.60 (.52)
19	.82	.09	.86/-1.4	.89/-1.1	.53 (.57)
15	.62	.09	.83/-1.7	.82/-1.8	.59 (.56)
4	.02	.10	.81/-1.8	.81/-1.8	.55 (.54)
3	-.36	.10	.78/-2.0	.76/-2.3	.58 (.52)
5	.00	.10	.76/-2.3	.75/-2.4	.60 (.54)

Next, the summary of category structure and observed average of endorsements was investigated. The observed frequencies and percentages for each possible response category revealed higher endorsement for “Neither Confident Nor Unconfident” at 30% of all responses and “Mostly Unconfident” at 12% of responses. The “Completely Unconfident” category only accounted for 5% of responses, while “Mostly Confident” had 38% and “Completely Confident” had 14%. This pattern of endorsements supports the positive cluster of persons above the mean on the item-person map, indicating slightly above average assessment confidence.

Infit and outfit MNSQ values appeared normal and did not markedly exceed 1.30. Also, the Andrich thresholds (i.e., difficulty between endorsing each response option) displayed a monotonically increasing average. The greatest increase was between “Mostly Confident” (1.37) and “Completely Confident” (3.47). The large logit gap between these categories demonstrated the relative ease in selecting between the “Mostly Confident” and “Completely Confident” categories compared to other adjacent categories. As shown in Figure 15 below, the lower three Likert categories have closer and less defined crests between “Completely Unconfident,” “Mostly Unconfident,” and “Neither Confident nor Unconfident.”



```
     CATEGORY PROBABILITIES: MODES - Andrich thresholds at intersections
    -+-------+-------+-------+-------+-------+-------+-------+-
1.0 +                                                         +
    |                                                         |
    |                                                         |
    |                                                       44|
 .8 +0                                                    44  +
    | 00                                                44    |
    |   00                                            44      |
    |     0                                         44        |
 .6 +      0                                       4          +
    |       00                      333333333    44           |
 .5 +         0                   33         33 4             +
    |          0        2222222 33             *3             |
 .4 +           0     22       *2            44  33           +
    |        1111*11**       33  22         4      33         |
    |     111     0*  111   3      22     44         33       |
    |   11       22 0    1*3         22 44             33     |
 .2 +111       22    0   3 11         4*2                333  +
    |        22       0*3    11     44   222                33|
    |      22       333 00     11*44        222               |
    |222222     3333      00***44 11111        222222         |
 .0 +***********444444444444   00000000***********************+
    -+-------+-------+-------+-------+-------+-------+-------+-
    -3      -2      -1       0       1       2       3       4
     PERSON [MINUS] ITEM MEASURE
```

Figure 15. Pilot Sample 13-Item Component 2 Confidence Category Probabilities (*N* = 165). This figure indicates the probability of endorsement for each of the Likert scale confidence categories, starting with “Completely Unconfident” on the left and ending with “Complete Confident” on the right. The strength of the peak between each category indicates the discrimination between response levels.

Table 32

Pilot Sample 13-Item Component 2 Confidence Measure Thresholds and Fit Table (N = 13)

Category Label	Observed Count	Observed Average	Sample Expected	Infit MNSQ	Outfit MNSQ	Andrich Threshold	Category Measure
0	112 (5%)	-.91	-1.04	1.14	1.19	None	(-2.94)
1	254 (12%)	-.30	-.40	1.11	1.25	-1.53	-1.44
2	652 (30%)	.11	.20	.97	.95	-1.04	-.23
3	822 (38%)	.86	.87	.97	.95	.29	1.37
4	305 (14%)	1.86	1.75	.89	.92	2.27	(3.47)

Component item characteristics summary. The Item Characteristics table below provides a broad overview of both general content and psychometric characteristics (i.e., reliability and validity) of the items on the modified CALI. Rasch PCA analyses compute general "Components" comprised of items that are related, and do not present set "Factors" of items that represent latent variables (Linacre, 1998). Brentari and Golia (2007) and Brentari, Golia, and Manisera (2007) describe the concept of dimensionality as a continuum and therefore even though statistical information may suggest multidimensionality, the dimensions exist on a continuum. The broad content characteristics present in this PCA continuum are: Applied, Methods, and Grading. These refer to the content of the question involving the application of knowledge versus declaratively recalling information, choosing an assessment strategy, and the inclusion of language relative to grading and scoring student performance. These general content characteristics were determined by a content analysis of the questions related to the themes of the *Standards* with which the measure aligned. The psychometric characteristics include length of question stem, use of a vignette, and item difficulty. The criterion for a lengthy item was 200 characters or less (i.e., without counting spaces). Vignettes are questions with stems that develop context. Lastly, difficulty was reported as the Rasch logit value which typically ranges from -3 to 3 with negative values indicating ease of endorsement and higher positive values with increased difficulty.

Table 33

Component Item Characteristics Summary

Item	Content Characteristics			Psychometric Characteristics		
	Applied	Methods	Grading	Lengthy	Vignette	Difficulty
1. What is the most important consideration in choosing a method for assessing student achievement?	N	Y	N	N	N	-.09
2. When scores from a standardized test are said to be “reliable,” what does it imply?	N	N	N	N	N	.35
3. Mrs. Bruce wished to assess her students' understanding of the method of problem solving she had been teaching. Which assessment strategy below would be most valid?	Y	Y	N	N	Y	-.62
4. What is the most effective use a teacher can make of an assessment that requires students to show their work (e.g., the way they arrived at a solution to a problem or the logic used to arrive at a conclusion)?	N	Y	N	N	N	.13
5. Ms. Green, the principal, was evaluating the teaching performance of Mr. Williams, the fourth-grade teacher. One of the things Ms. Green wanted to learn was if the students were being encouraged to use higher order thinking skills in the class. What documentation would be the most valid to help Ms. Green to make this decision?	Y	N	N	Y	Y	.68
6. Ms. Gregory wants to assess her students' skills in organizing ideas rather than just repeating facts. Which words should she use in formulating essay exercises to achieve this goal?	Y	Y	Y	N	Y	.50
7. Mr. Woodruff wanted his students to appreciate the literary works of Edgar Allen Poe. Which of his test items shown below will best measure his instructional goal?	Y	N	N	N	Y	-.45

8. Several students in Ms. Atwell's class received low scores on her end-of-unit test covering multi-step story problems in mathematics. She wanted to know which students were having similar problems so she could group them for instruction. Which assessment strategy would be best for her to use for grouping students?	Y	Y	N	Y	Y	-.21
9. Many teachers score classroom tests using a 100-point percent correct scale. In general, what does a student's score of 90 on such a scale mean?	N	N	Y	N	N	.08
10. Students in Mr. Jakman's science class are required to develop a model of the solar system as part of their end-of-unit grade. Which scoring procedure below will maximize the objectivity of assessing these student projects?	Y	N	Y	N	Y	-.42
11. At the close of the first month of school, Mrs. Friend gives her fifth grade students a test she developed in social studies. Her test is modeled after a standardized social studies test. It presents passages and then asks questions related to understanding and problem definition. When the test was scored, she noticed that two of her students – who had been performing well in their class assignments – scored much lower than other students. Which of the following types of additional information which would be most helpful in interpreting the results of this test?	Y	N	Y	Y	Y	-.17
12. When the directions indicate each section of a standardized test is timed separately, which of the following is acceptable test-taking behavior?	N	N	N	N	N	-.54
13. Ms. Camp is starting a new semester with a factoring unit in her Algebra I class. Before beginning the unit, she gives	Y	N	N	Y	Y	.66

her students a test on the commutative, associative, and distributive properties of addition and multiplication. Which of the following is the most likely reason she gives this test to her students?						
14. To evaluate the effectiveness of the mathematics program for her gifted first graders, Ms. Allen gave them a standardized mathematics test normed for third graders. To decide how well her students performed, Ms. Allen compared her students' scores to those of the third-grade norm group. Why is this an incorrect application of standardized test norms?	Y	N	Y	Y	Y	.41
15. A teacher gave three tests during a grading period and she wants to weight them all equally when assigning grades. The goal of the grading program is to rank order students on achievement. In order to achieve this goal, which of the following should be closest to equal?	Y	N	Y	Y	Y	.53
16. When a parent asks a teacher to explain the basis for his or her child's grade, the teacher should…	N	N	Y	N	N	-.32
17. Which of the following grading practices results in a grade that least reflects students' achievement?	N	N	Y	N	N	.08
18. During the most recent grading period, Ms. Johnson graded no homework and gave only one end-of-unit test. Grades were assigned only on the basis of the test. Which of the following is the major criticism regarding how she assigned the grades?	Y	N	Y	Y	Y	.10
19. In a routine conference with Mary's parents, Mrs. Estes observed that Mary's scores on the state assessment program's quantitative reasoning tests indicate Mary is performing better in mathematics concepts than in mathematics computation. This probably means that…	Y	N	Y	Y	Y	-.22
20. Mr. Klein bases his students' grades mostly on graded	Y	N	Y	Y	Y	-.36

homework and tests. Mr. Kaplan bases his students' grades mostly on his observation of the students during class. A major difference in these two assessment strategies for assigning grades can best be summarized as a difference in…						
21. John scored at the 60th percentile on a mathematics concepts test and scored at the 57th percentile on a test of reading comprehension. If the percentile bands for each test are five percentile ranks wide, what should John's teacher do in light of these test results?	Y	N	Y	Y	Y	.43
22. In some states testing companies are required to release items from prior versions of a test to anyone who requests them. Such requirements are known as	N	N	N	N	N	-.21
23. Mrs. Brown wants to let her students know how they did on their test as quickly as possible. She tells her students that their scored tests will be on a chair outside of her room immediately after school. The students may come by and pick out their graded test from among the other tests for their class. What is wrong with Mrs. Brown's action?	Y	N	Y	Y	Y	.11
24. A state uses its statewide testing program as a basis for distributing resources to school systems. To establish an equitable distribution plan, the criterion set by the State Board of Education provides additional resources to every school system with student achievement test scores above the state average. Which cliché best describes the likely outcome of this regulation?	Y	N	N	Y	Y	-.44
25. Mrs. Overton was concerned that her students would not do well on the State Assessment Program to be administered in the Spring. She got a copy of the standardized test form that was going to be used. She did	Y	N	N	Y	Y	-.62

each of the following activities to help increase scores.
Which activity was unethical?

Research Question 2

The second research question addressed in the current study stated: "What is the impact of assessment confidence on the relationship between pre-service teachers' assessment literacy and performance assessment scores?" This analysis required a second descriptive investigation of the sample and its demographics, as the sample used in the second phase of the study was different from the pilot sample. Moderated Multiple Regression Analyses were then employed to assess the relationship between assessment knowledge and assessment confidence, as measured by the CALI and edTPA performance.

Descriptives

Sample descriptive information ($N = 112$) from the second phase of data collection (i.e., mostly 4th-year students and a very small number of graduate-level teacher education students across the same variety of programs as the pilot) contained 20 males (17.9%) and 92 females (82.1%). This is consistent with current undergraduate enrollment trends as well as the general population of teachers across the nation (Peter et al., 2005). Of these participants, the average age was 23.28 ($SD = 2.77$; $Mdn = 22.00$, IQR = 2) with a range of 20 to 35 years old. According to university credit requirements, 102 participants (91.1%) had senior status. The remaining ten participants had graduate-level student status. The majority of participants were White/Caucasian ($n = 102$, 91.1%), with 8.9% of participants reporting a minority race.

There were 35 (31.3%) first-generation college students in this sample. Thirty-point four percent of the participants were in an ECED program, 12.5% were in MCED, 25.9% were in AYA, and 31.2% were in various teaching preparation programs such as Foreign Language, English as a Second Language, Health or Physical Education, Art, and Music. Additionally, the average self-reported cumulative GPA was 3.62 (*SD* = .27). Regarding parent education levels, 24.1% of the participants' mothers and 28.6% of their fathers received a High School diploma or GED, and 35.7% of their mothers and 27.7% of their fathers have a Bachelor's Degree. The demographic variable group proportions in this sample are comparable to the pilot sample.

The following paragraph summarizes some additional descriptive information related to student course work (i.e., not demographics) that was not included in the below table. In relation to student experience with assessment in their coursework, 26.8% (*n* = 30) reported taking a course solely focused on assessment, with 5.4% (*n* = 6) taking more than one assessment-only course.

In this sample, 16.1% (*n* = 18) had attended a workshop with an assessment-only focus, and 7.2% (*n* = 8) attended more than one workshop. Courses that have at least one assessment lesson/unit were attended by 90.2% (*n* = 101) of the participants, with 83.9% (*n* = 94) having attended more than one. Workshops that had an assessment component were attended by 34.8% (*n* = 39) of the participants and 17.9% (*n* = 20) attended more than one. Lastly, the majority of participants felt they were "Somewhat Prepared" (*n* = 41, 36.6%) or "Very Prepared" (*n* = 64, 57.1%) to be a teacher based on their training. These students similarly perceived that they were prepared to assess student learning with 51 (45.5%) participants indicating that they were "Somewhat Prepared" and 42 (37.5%) students feeling "Very Prepared."

Finally, the second phase sample, who completed the 25-item modified CALI, had an average score of 14.54 (*SD* = 3.77), and their average Assessment Confidence score was 2.67 (*SD* = .55; *Mdn* = 2.74, IQR = .83). The internal consistency reliability for the 25-item CALI was .649 (KR-20) and for the 25-item confidence measure was .923 (Cronbach's α). Table 34 (see below) summarizes the second phase sample demographic and other descriptive variables from the preceding paragraphs. Total CALI and Assessment Confidence score descriptive statistics were reported for each categorical demographic and descriptive variable in order to provide evidence of group equivalence prior to conducting the confirmatory phase of measure development.

Table 34

Second Phase Sample Variable Descriptive Statistics: Classroom Assessment Literacy Inventory (CALI) Total (25 Item) Scores and Assessment Confidence (Average) (25 Item) Scores (N = 112)

Variable		CALI Total			Assessment Confidence Total		
	n	*M/Mdn*	*SD*/IQR	Min/Max	*M/Mdn*	*SD*/IQR	Min/Max
Gender							
Male	20	13.30	4.05	5/19	2.62	.54	.56/3.52
Female	92	14.82	3.67	6/21	2.68/2.74*	.56/.81	1.36/3.40
Age	112	23.28/22.00*	2.77/2.00	20/35	--	--	--
Race							
White/Caucasian	102	14.44	3.78	5/21	2.66/2.72*	.52/.82	.80/3.52
Other	10	15.60	3.72	7/20	2.75/3.00*	.83/.70	.56/3.52
1st Generation College Student							
Yes	35	14.89	3.91	5/21	2.81	.43	1.96/3.44
No	77	14.39	3.72	7/20	2.60/2.64*	.59/.76	.56/3.52
GPA	112	3.62	.27	2.96/4.00	--	--	--
Program							
ECED	34	14.94	3.41	6/20	2.71/2.88*	.59/.62	.80/3.52
MCED	14	15.43	3.78	9/21	2.66	.59	1.36/3.32
AYA	29	14.93/16.00*	3.53/4.00	5/19	2.83/2.96*	.54/.52	.56/3.40
Other	35	13.49	4.19	7/21	2.49	.55	1.72/3.52
Year							
Senior	102	14.50	3.82	5/21	2.67/2.74*	.55/.81	.56/3.52
Other	10	15.00	3.40	9/19	2.64	.60	1.72/3.40
Mother’s Education							
HSD/GED or Less	27	15.11	4.17	5/21	2.76	.41	1.96/3.32
Some College/Associate/Tech	29	14.83	3.07	9/20	2.77/2.84*	.56/.62	.80/3.52
Bachelor’s Degree	40	13.70	3.66	7/19	2.55/2.64*	.57/.82	.56/3.40
Master’s/Doctoral/Professional	16	15.19	4.40	7/20	2.62	.69	1.04/3.52

Father’s Education							
HSD/GED or Less	32	14.41	4.10	5/21	2.69/2.92*	.58/.99	.80/3.44
Some College/Associate/Tech	34	14.50	3.96	7/21	2.61/2.72*	.57/.85	.56/3.40
Bachelor’s Degree	31	14.51	3.26	8/20	2.69	.58	1.04/3.52
Master’s/Doctoral/Professional	15	15.00	3.93	7/20	2.70	.42	1.88/3.20
Course(s) with Assessment							
Yes Took Course(s)	101	14.58	3.54	5/21	2.66/2.76*	.57/.80	.56/3.52
No Did Not	11	14.18	3.81	7/19	2.73	.40	2.08/3.20
Assessment-Specific Course							
Yes Took Course	30	13.97	4.72	5/21	2.57	.62	.56/3.52
No Did Not	82	14.76	3.37	6/21	2.70/2.84*	.53/.67	.80/3.52
Student Teaching Experience							
Yes	107	14.62	3.70	6/21	2.66/2.72*	.55/.80	.56/3.52
No	5	13.00	5.39	5/18	2.84/3.00*	.52/.80	1.96/3.24
CALI Total	112	14.54	3.77	5/21	--	--	--
Assessment Confidence Total	112	--	--	--	2.67/2.74*	.55/.83	.56/3.52

Note. Groups and continuous variables denoted with asterisks next to the values in the *M/Mdn* columns indicate non-normal distributions.

As with the pilot sample, relationships between these demographics variables and the Total CALI and Confidence scores were examined to provide evidence of group equivalence (see Table 35). The measurement levels of the variables and the CALI and Confidence score skewness dictated the statistical tests used (i.e., parametric or nonparametric). Pearson or Spearman correlations were used to examine relationships between two continuous or ordinal variables. Independent *t*-Tests and One-Way ANOVAs (i.e., or their nonparametric equivalents) were selected to investigate CALI and Confidence score differences between groups with two or three or more levels of the categorical variable. The only variables that had statistically significant (and positive) relationships with the Total CALI score were GPA ($r = .275, p = .003$) and Total (average) Assessment Confidence ($r_s = .593, p < .001$). There were also significant relationships between Assessment Confidence and 1st generation college student status ($p = .047$) and program ($p = .016$).

Table 35

Relationships between Second Phase Sample Variables and Classroom Assessment Literacy Inventory (CALI) Total Scores (25 Item) and Assessment Confidence Total Scores (N = 112)

Variable	CALI Total		Assessment Confidence Total	
	Statistical Test	*p*	Statistical Test	*p*
Gender	$t(110) = 1.641$	.104	$U = 852.500, Z = -.513$	.608
Age	$r_s = -.055$	.563	$r_s = -.014$	.881
Race	$t(110) = .927$	.356	$U = 389.500, Z = -1.230$	.219
1st Generation College Student	$t(110) = -.644$	.521	$U = 1031.00, Z = -1.988$	$.047^{*}$
GPA	$r = .275$	$.003^{**}$	$r_s = .126$	.187
Program	$H(3) = 3.224$	.358	$H(3) = 10.353$	$.016^{*}$

Year	$t(110) = -.399$	.691	$U = 486.500, Z = -2.40$	.810
Mother's Education	$F(3,108) = 1.084$	.359	$H(3) = 4.189$	.242
Father's Education	$F(3,108) = .087$	.967	$H(3) = .621$	.892
Course(s) with Assessment	$t(110) = -.335$	.738	$U = 545.00, Z = -.103$	.918
Assessment-Specific Course	$t(40.334) = .842$	.405	$U = 1083.00, Z = -$	.335
Student Teaching	$t(110) = -.937$	.351	.963	.405
Experience	$r_s = .593$	$<.001^{***}$	$U = 208.50, Z = -.832$	--
Assessment Confidence			--	
Total				

Note. $^{*}p < .05$, $^{**}p <.01$, $^{***}p < .001$.

Confirmatory Factor Analysis (CFA)

A CFA was performed in order to examine the internal structure of the modified CALI. Before conducting the CFA, univariate and multivariate log likelihood outliers were examined using Maximum Likelihood (ML) estimation by treating the indicators as continuous. No univariate or multivariate outliers were identified and therefore this analysis consisted of the full sample. Additionally, there were no missing values in the modified CALI (Muthen, 1983b).

Model estimation. The two estimation methods for dichotomous data are Unweighted Least Squares (ULS) and Weighted Least Squares (WLS; Brown, 2012; Yang-Wallentin et al., 2010). WLS can also be used for ordinal and dichotomous data, but still requires large sample sizes, and was therefore not used in this study. ULS and DWLS are similar to WLS but differ by using a weight matrix under the fit function. DWLS uses a weight matrix, which only contains the diagonal elements of the asymptotic covariance matrix, while ULS uses the identity matrix as its weight matrix (Schumacker & Lomax, 2010). Previous research has shown that ULS outperforms DWLS and gives more precise estimation by means of less bias and smaller standard errors than DWLS (Forero et al., 2009; Rigdon & Ferguson, 1991). ULS is also recommended when a polychoric correlation matrix is used because it considers the weight matrix and non-convergence (Babakus et al., 1987). For this reason, ULS was used for all CFA analyses of dichotomous data representing the item scores.

Parceling. Unweighted Least Squares (ULS) was selected for model estimation based on small sample size and previous research (Savalei, Bonnett, & Bentler, 2015). However, the dichotomous data coupled with the likelihood of producing a non-positive definite correlation matrix necessitated using procedures to create ordinal variables using parceling. Parceling is a measurement aggregation technique often used in CFA and Structural Equation Modeling (SEM) to sum or average two or more items to create an ordinal or continuous variable (Little, Cunningham, & Shahar, 2002). Item parceling is best used when data are nonnormally distributed and/or coarsely categorized, which are two conditions that violate the assumption of several CFA estimation methods (Bandalos, 2002). Ultimately, these concerns manifest in CFA model fit indices via Chi-Square values and standard errors (Bandalos). However, parceling is a controversial practice that many psychometricians differ on philosophically. The center of the anti-parceling philosophy believes that by creating parcels the data do not represent the “real” model as the structure becomes manufactured. This debate will be further presented in Chapter 5.

Conversely, many psychometricians believe item parceling is a pragmatic approach to assessing underlying latent variables. When the content of a measure is of interest, parceling can successfully be employed as a component of CFA (Lawrence & Dorans, 1987). Parceling practices are not advisable for analyses focused on item-level information (i.e., standardized tests). Therefore, parceling was not employed other than in this CFA, which was focused more on the content of the items (Little, Rhemtulla, Gibson, & Schoemann, 2013). Since the content of the CALI is of primary focus, creating item parcels, as opposed to using item-level data, increases the reliability, communality, and the ratio of common-to-unique factor variance (Little et al.). Through parceling, the data are evenly distributed which allows for an interpretable analysis. Additionally, item parceling can mitigate matrix-level concerns possible in CFA (Lawrence & Dorans) and reduce the number of parameters required to define a construct, as item-level data requires more parameters than parcels (Little, Cunningham, & Shahar, 2002). This final point of emphasis is extremely relevant with small sample sizes which better fit models with fewer parameters.

There are several practices for creating item parcels, varying from random assignment, to using item difficulty levels and item content. Prior to creating parcels, the dimensionality of items to be parceled must be determined (Orcan, 2013). In the case of the modified CALI, dimensionality was achieved in the creation of two components produced in the pilot Rasch PCA – Component 1 consisting of 12 items and Component 2 consisting of 13 items. Parcels were then created for the 25-item modified CALI by first running a Rasch Analysis on both of the previously defined components. An item-to-construct approach was used to create parcels within each of the two components. The item-to-construct balance, outlined by Little, Cunningham, and Sharhar (2002), equally balances items within each parcel based on their difficulty level and content. This means that all difficult items in Component 1, for example, could not be in the same parcel.

Additionally, random assignment was used to control for confounds such as item type (e.g., vignettes versus no vignettes). While parcels were balanced on difficulty, content, and item type, they were also randomly assigned to each parcel until this balance was achieved.

Component 1 parcels. A Rasch analysis was conducted on all 12 items in Component 1 (KR-20 = .682) and the component was found to be unidimensional as no more than one component was present (*M* = .00, *SD* = .61; RMSE = .25, *SD* = .56; Separation = 2.26, Item Reliability = .84). Information from the Component 1 Rasch Analysis created four parcels, each with three items ranging from logit values of -.88 to 1.49 on the vertical scale. Since the logit values provide an indication of item difficulty related to this sample, they were used as a starting point for creating parcels. The four most difficult items (i.e., Items 25, 14, 17, and 6) ranged from .31 to 1.49 logits, the four central items (i.e., Items 10, 24, 9, and 3) from -.28 to .00, and the easiest four items (i.e., Items 18, 23, 12, and 1) from -.88 to -.34 logits. All items were then assessed according to their content as this dimension consisted of content points including scoring and grading student performance, and testing/assessment ethics. The final parcel structure was: Component 1 Parcel 1 Items 25, 9, and 16; Component 1 Parcel 2 Items 14, 10, and 12; Component 1 Parcel 3 Items 6, 24, and 18; Component 1 Parcel 4 Items 17, 13, and 23.

Component 2 parcels. A Rasch Analysis was conducted on all 13 items in Component 2 (KR-20 = .322) and the component was found to be unidimensional as no more than one dimension was present (*M* = .00, *SD* = 1.81; RMSE = .29, *SD* = 1.79;

Separation = 6.15, Item Reliability = .97). Information from the Component 2 Rasch Analysis created four parcels, three of which had three items and the fourth with four items. Overall, the items ranged in logit values from -3.44 to 3.06 on the vertical scale. Since the logit values provide an indication of item difficulty related to this sample, they were used as a starting point for creating parcels. The four most difficult items (i.e., Items 22, 21, 15, and 8) ranged from .98 to 3.06 logits, the five central items (i.e., Items 5, 2, 19, 11, and 20) from -.24 to .66, and the easiest four items (i.e., Items 4, 7, 3, and 1) from -3.44 to -1.11 logits. All items were then assessed according to their content as this component consisted of content points including choosing assessment methods and strategies and communicating assessment results. The final parcel structure was: Component 2 Parcel 1 Items 15, 2, and 1; Component 2 Parcel 2 Items 22, 19, and 3; Component 2 Parcel 3 Items 21, 11, and 7; Component 2 Parcel 4 Items 8, 5, 20, and 4.

Table 36

Summary of Parcel Information (N = 8)

Component #	Parcel #	Item #s
1	1	25, 9, 16
1	2	14, 10, 12
1	3	6, 24, 18
1	4	17, 13, 23
2	1	15, 2, 1
2	2	22, 19, 3
2	3	21, 11, 7
2	4	8, 5, 20, 4

Model identification. The initial model included the eight parcels, introduced above (i.e., eight observed variables), created from the modified CALI data. There were four parcels for each of the two components and loaded on only one of the two possible dimensions (i.e., Component 1 or Component 2), which were identified in the PCA exploratory analysis. A summary of items/parcels and the dimensions that they load on are included in Table 36. All the measurement errors were presumed to be unsystematic (i.e., there are no correlated measurement errors for any pairs of parcels; Brown, 2015).

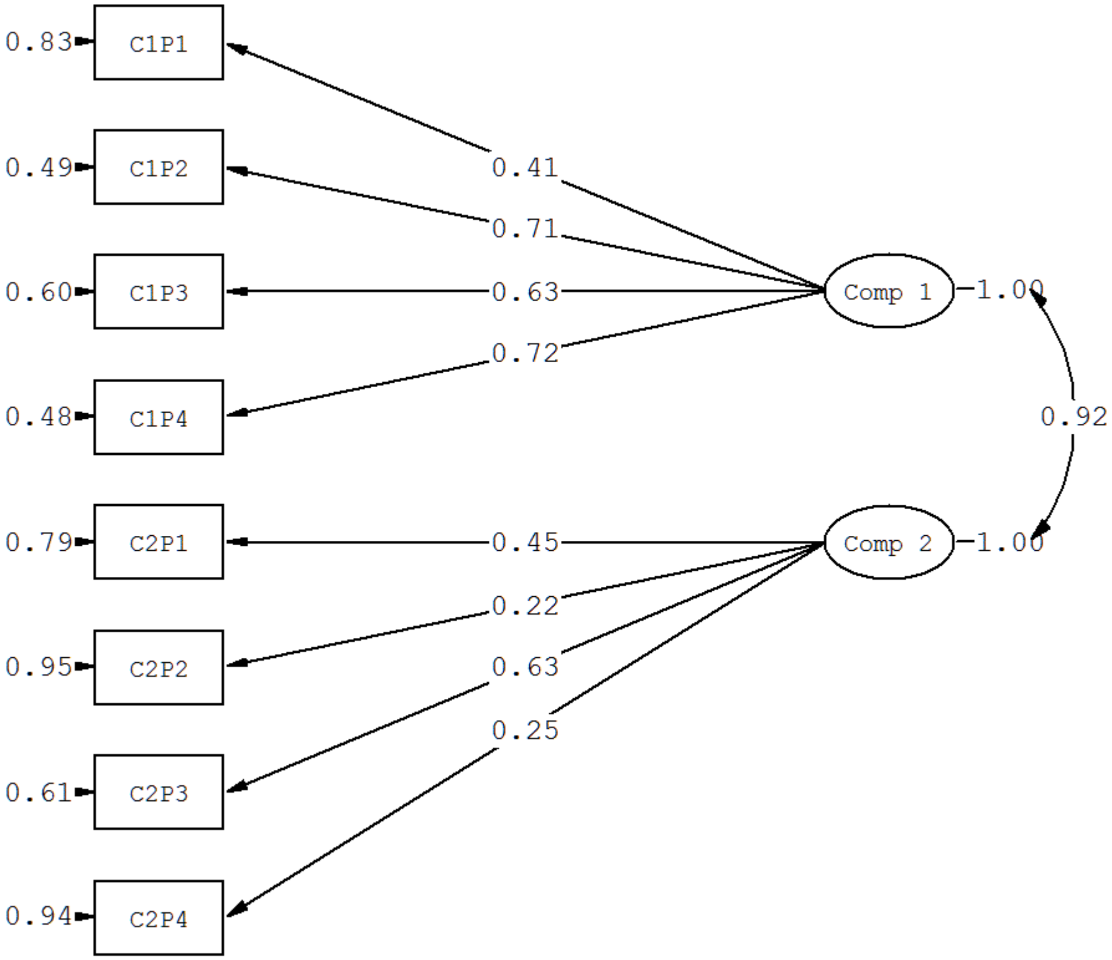


Figure 16. CFA Graphical Representation. The figure is the CFA initial model for the modified CALI. The standardized parameter estimates for the factor structure of the modified CALI are listed in the following table. Rectangles represent the 8 parcel scores (i.e., observed variables) and the ovals represent the two latent factors, which were hypothesized in the PCA analysis.

There were eight factor loadings, eight measurement errors, and one factor correlation. Because the distinct values (i.e., unique values) in the matrix *S* (36) are greater than the total number of free parameters (19), this model is considered over-identified (i.e., there is more than one way of estimating parameters; Schumacker & Lomax, 2010). The hypothesized model presented conflicting data fit across five different fit indices ($\chi^2 = 50.911$, $df = 19$, $p = .001$; GFI = .963; AGFI = .929; RMSEA = .113; RMR = .081; SRMR = .081). For example, the GFI and AGFI values were greater than or close to .95, which is an indicator of good fit. However, the Chi-Square value, the RMSEA value, and the SRMR value were greater than .05 which indicated poor fit (Schreiber, Nora, Stage, Barlow, & King, 2006). It should be noted that the Brown's (1984) ADF Chi-Square value should be used with ULS models as it provides the most accurate value (Schumacker & Lomax, 2010). Additionally, in each model presented in this analysis one of the paths in each factor was fixed to a value of one. Fixing the path to one allows shows the relationship between the latent variable and the observed variable and allows for a determination of the variance of the latent variable.

The standardized loadings represent the correlation between each observed variable and the corresponding factor (Schumacker & Lomax, 2010). Table 37 summarizes the factor loadings and their significance values. While some precise factor loading values have been suggested by Comrey and Lee (1992), Guadagnoli and Velicer (1988) regard a factor as reliable if it has four or more loadings of at least .6, regardless of sample size. MacCallum, Widman, Zhang and Hong (1999) agree with the loadings of at least .6 to justify performing a factor analysis with small sample sizes. Lastly, Stevens

(1992) suggests using a cut-off of .4, irrespective of sample size, for interpretative purposes. Given this information and the reported factor loadings, as summarized in Table 37, all parcels had moderate to strong factor loading values and all loadings were significant. The amount of variance in each observed variable was accounted for and ranged from 5.0% to 52.0%.

Table 37

Summary of Factor Loading Information (N = 8)

Parcel	Factor Loading	p	R^2
Component 1 Parcel 1	.41	.000	.171
Component 1 Parcel 2	.71	.006	.510
Component 1 Parcel 3	.63	.000	.399
Component 1 Parcel 4	.72	.000	.522
Component 2 Parcel 1	.45	.000	.206
Component 2 Parcel 2	.22	.003	.049
Component 2 Parcel 3	.63	.001	.391
Component 2 Parcel 4	.25	.001	.064

Note. Factor loading values are standardized estimates.

Model modification. The modification indices recommended adding error covariances among observed variables. Adding an error covariance allows for the variance between two variables that show similar behavior to correlate. This process continued until adding more parameters did not produce a significant improvement in model fit. There were two covariances added in the modification stage. The first modification suggested was adding an error covariance between Component 2 Parcel 2 and Component 2 Parcel 1. Since these parcels loaded on the same factor throughout these analyses and the factor is representative of a general content domain (e.g., content points including choosing assessment methods and strategies and communicating assessment results) the covariance was added. The second modification added an error covariance between Component 2 Parcel 1 and Component 1 Parcel 4. This covariance went across Components 1 and 2 and required investigation of item content in both parcels. Component 2 Parcel 1 Item 15 utilized content about creating assessments (Component 2) but also had aspects of scoring (Component 1). Additionally, Component 2 Parcel 1 Item 2 focused on understanding assessment results, which is related to using assessment results (Component 1 Parcel 4 Item 13). Thus, the modification was warranted.

The final model had acceptable fit ($\chi^2 = 23.774$, $df = 17$, $p = .1257$; GFI = .980; AGFI = .957; RMSEA = .062; RMR = .059; SRMR = .059). A comparison of all model fit indices is presented in Table 38. The final model (i.e., including the error covariances) is illustrated in Figure 17. All standardized loadings were moderate to large and statistically significant ($p < .01$ for all).

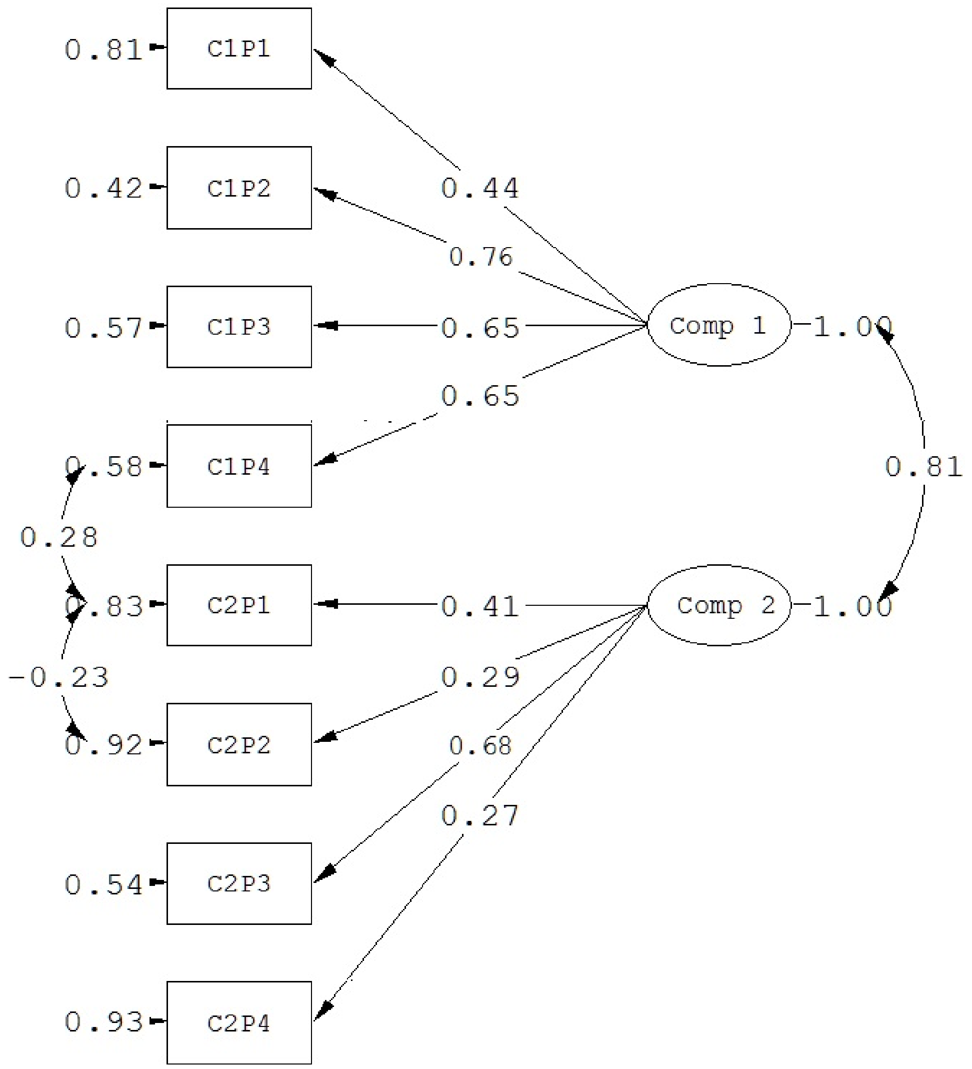


Figure 17. CFA Final Model Graphical Representation. The figure is the CFA final model for the modified CALI with standardized parameter estimates.

Table 38

Model Modification and Fit Index Comparison Information

Model	Chi-Square	GFI	AGFI	RMSEA	RMR	SRMR
Initial Model	50.911	.963	.929	.113	.081	.081
Second Model	39.185	.971	.942	.099	.070	.071
Final Model	23.774	.980	.957	.062	.059	.059

Note. The second model had an error covariance between Component 2 Parcel 2 and Component 2 Parcel 1. The final model consisted of this error covariance and added an error covariance between Component 2 Parcel 1 and Component 1 Parcel 4.

Summary of All Psychometric Results

There are two key results from the Rasch Analyses, Rasch PCA, and the CFA in relation to the dimensionality of the modified CALI. The Rasch analysis of the 25-item modified CALI produced good psychometric properties(i.e., reliability and validity) . Therefore, this analysis supported a unidimensional 25-item measure of assessment knowledge and confidence. However, the subsequent Rasch PCA provided evidence of a potential second dimension within the 25-item measure. The CFA established the possibility of a second component using a second sample. It must be noted that these two components exist on the same continuum of assessment literacy which after brief content review cover broad domains of applied knowledge, methods, and grading. However, as these results were the first to demonstrate these potential broader categorizations of assessment knowledge as measured by the CALI, more replicative support is needed. Given that there is psychometric support for both internal structures of the CALI, the following analyses (i.e., the remaining Research Question 2 analyses) used the unidimensional 25-item measure and the two-component measure.

Table 39

Component 1 Knowledge and Confidence: Second Phase Sample Variable Descriptive Statistics for (12 Item) Classroom Assessment Literacy Inventory (CALI) Scores and Assessment Confidence Scores (N = 112)

Variable		CALI Component 1			Confidence Component 1		
	n	*M/Mdn*	*SD*/IQR	Min/Max	*M/Mdn*	*SD*/IQR	Min/Max
Gender							
Male	20	6.95	2.82	2/10	2.77	.61	1.67/3.67
Female	92	8.42/9.00*	2.47/3.00	2/12	2.83/3.00*	.66/.83	.58/3.75
Age	112	23.28/22.00*	2.77/3.00	20/35	--	--	--
Race							
White/Caucasian	102	8.07	2.59	2/12	2.81	.62	.92/3.75
Other	10	9.10/9.50*	2.46/3.00	3/11	2.89/2.92*	.89/.91	.58/3.58
1st Generation College Student							
Yes	35	8.29	2.16	3/11	2.99	.56	1.67/3.75
No	77	8.10	2.77	2/12	2.74	.67	.58/3.75
GPA	112	3.62	.27	2.96/4.00	--	--	--
Program							
ECED	34	8.71/10.00*	2.42/3.00	2/11	2.85/3.00*	.66/.70	.92/3.67
MCED	14	8.71	2.55	4/12	2.81	.58	1.67/3.50
AYA	29	8.41/9.00*	2.31/2.00	2/11	3.02/3.33*	.64/.75	.58/3.75
Other	35	7.20	2.79	2/11	2.63	.63	1.67/3.75
Year							
Senior	102	8.14	2.63	2/12	2.82/3.00*	.64/.85	.58/3.75
Other	10	8.40	2.17	4/11	2.79	.74	1.67/3.67
Mother’s Education							
HSD/GED or Less	27	8.48/10.00*	2.49/2.00	2/11	2.95	.57	1.67/3.75
Some College/Associate/Tech	29	8.38	1.86	4/11	2.91/3.00*	.61/.66	.92/3.75
Bachelor’s Degree	40	7.57	3.01	2/12	2.69	.65	.58/3.67
	16	8.69	2.75	2/11	2.75	.79	1.08/3.67

Master’s/Doctoral/Professional							
Father’s Education							
HSD/GED or Less	32	8.06	2.41	2/11	2.82	.70	.92/3.75
Some College/Associate/Tech	34	8.15	2.86	2/12	2.76/2.96*	.65/.85	.58/3.42
Bachelor’s Degree	31	8.19	2.48	2/12	2.86	.66	.66/1.08
	15	8.33	2.74	3/11	2.88	.55	.55/1.92
Master’s/Doctoral/Professional							
Course(s) with Assessment							
Yes Took Course(s)	101	8.19	2.58	2/11	2.81	.66	.58/3.75
No Did Not	11	7.82	2.75	2/12	2.88	.54	2.00/3.67
Assessment-Specific Course							
Yes Took Course	30	7.57	3.10	2/12	2.71	.74	.58/3.75
No Did Not	82	8.38	2.35	2/12	2.86	.61	.92/3.75
Student Teaching Experience							
Yes	107	8.19	2.58	2/12	2.81	.64	.58/3.75
No	5	7.40	2.97	3/10	2.97	.77	1.67/3.67
CALI Component 1	112	8.16/9.00*	2.58/3.00	2/10	--	--	--
Confidence Component 1	112	--	--	--	2.82/3.00*	.65/.89	.58/3.75

Note. Groups and continuous variables denoted with asterisks next to the values in the *M/Mdn* columns indicate non-normal distributions.

Table 40

Component 2 Knowledge and Confidence: Second Phase Sample Variable Descriptive Statistics for (13 Item) Classroom Assessment Literacy Inventory (CALI) Scores and Assessment Confidence Scores (N = 112)

Variable		CALI Component 2			Confidence Component 2		
	n	*M*/*Mdn*	*SD*/IQR	Min/Max	*M*/*Mdn*	*SD*/IQR	Min/Max
Gender							
Male	20	6.35	1.69	2/9	2.49	.51	.54/3.46
Female	92	6.39	1.82	2/11	2.53	.52	1.08/3.15
Age	112	23.28/22.00*	2.77/2.00	20/35	--	--	--
Race							
White/Caucasian	102	6.37	1.82	2/11	2.52/2.89*	.48/.56	.69/3.46
Other	10	6.50	1.51	4/9	2.62/2.58*	.82/.62	.54/3.46
1st Generation College Student							
Yes	35	6.60	2.12	2/11	2.65/2.77*	.41/.69	1.77/3.46
No	77	6.29	1.62	3/11	2.46	.55	.54/3.46
GPA	112	3.62	.27	2.96/4.00	--	--	--
Program							
ECED	34	6.24	1.72	2/9	2.59/2.67*	.57/.58	.69
MCED	14	6.71	1.86	4/11	2.53	.58	1.08
AYA	29	6.52	1.72	2/9	2.65/2.69*	.49/.38	.54
Other	35	6.29	1.92	3/10	2.36	.43	1.69
Year							
Senior	102	6.36	1.80	2/11	2.53/2.62*	.52/.69	.54/3.46
Other	10	6.60	1.71	3/9	2.50	.49	1.77/3.15
Mother's Education							

HSD/GED or Less	27	6.63	2.15	2/11	2.59	.37	1.77/3.08
Some College/Associate/Tech	29	6.45	1.68	4/10	2.63/2.69*	.55/.66	.69/3.46
Bachelor's Degree	40	6.13	1.47	4/9	2.42/2.54*	.54/.60	.54/3.15
Master's/Doctoral/Professional	16	6.50	2.09	3/9	2.49	.61	1.00/3.46
Father's Education							
HSD/GED or Less	32	6.35	2.15	2/11	2.57	.53	.69/3.46
Some College/Associate/Tech	34	6.35	1.67	3/10	2.48	.55	.45/3.38
Bachelor's Degree	31	6.32	1.59	3/10	2.54	.54	1.00/3.46
Master's/Doctoral/Professional	15	6.67	1.72	4/9	2.52	.37	1.85/3.00
Course(s) with Assessment							
Yes Took Course(s)	101	6.39	1.84	2/11	2.52	.53	.54/3.46
No Did Not	11	6.36	1.21	5/8	2.60/2.62*	.35/.64	2.00/3.08
Assessment-Specific Course							
Yes Took Course	30	6.40	2.13	2/10	2.45	.57	.54/3.31
No Did Not	82	6.38	1.66	2/11	2.55	.49	.59/3.46
Student Teaching Experience							
Yes	107	6.42	1.75	2/11	2.52	.53	.54/3.46
No	5	5.60	2.61	2/8	2.72/2.85*	.29/.47	2.23/3.00
CALI Component 2	112	6.38	1.78	2/9	--	--	--
Confidence Component 2	112	--	--	--	2.52/2.62*	.52/.69	.54/3.46

Note. Groups and continuous variables denoted with asterisks next to the values in the *M/Mdn* columns indicate non-normal distributions.

Table 41

Second Phase Sample Variables and Classroom Assessment Literacy Inventory (CALI) Component 1 Score and Confidence Component 1 Score (12 Item) Relationships (N = 112)

Variable	CALI Component 1		Confidence Component 1	
	Statistical Test	p	Statistical Test	p
Gender	$U = 613.00, Z = -2.366$	.018*	$U = 843.50, Z = -.582$	.560
Age	$r_s = -.081$	.397	$r_s = -.012$	.902
Race	$U = 371.00, Z = -1.439$	.150	$U = 429.50, Z = -.823$	.410
1st Generation	$t(110) = -.344$	.732	$t(110) = -1.881$	.063
GPA	$r = .304$	<.001***	$r = .136$	.152
Program	$H\ (3) = 7.603$	.055	$H(3) = 7.945$	.047*
Year	$t(110) = -.306$	.760	$U = 488.00, Z = -.225$	.822
Mother's Education	$H\ (3) = 3.147$	.370	$H(3) = 3.633$	.304
Father's Education	$F(3,108) = .039$	.990	$H(3) = .659$	.883
Course(s) with Assessment	$t(110) = .982$	.645	$t(110) = .277$	.782
Assessment-Specific Course	$t(110) = 1.480$	.142	$t(110) = 1.082$	.281
Student Teaching Experience	$t(110) = -.672$	.503	$t(110) = .521$	.604

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Table 42

Second Phase Sample Variables and Classroom Assessment Literacy Inventory (CALI) Component 2 Score and Confidence Component 2 Score (13 Item) Relationships (N = 112)

Variable	CALI Component 2		Confidence Component 2	
	Statistical Test	p	Statistical Test	p
Gender	$t(110) = .093$	.926	$t(110) = .325$	.746
Age	$r_s = -.005$	.958	$r_s = .001$	.991
Race	$t(110) = .214$	.831	$U = 349.00, Z = -1.64$	.100
1st Generation	$t(110) = -.862$	.391	$t(110) = -1.789$	.076
GPA	$r = .140$	.142	$r = .133$	.161
Program	$F(3,108) = .321$	.810	$H(3) = 10.595$	.014*
Year	$t(110) = -.399$	.691	$U = 483.00, Z = -.276$	.782
Mother's Education	$F(3,108) = .478$	.698	$H(3) = 3.555$	.314
Father's Education	$F(3,108) = .143$	.934	$F(3,108) = .164$	.921
Course(s) with Assessment	$t(110) = -.039$	.969	$U = 520.00, Z = -.348$	.728
Assessment-Specific Course	$t(110) = -.057$	.954	$t(110) = .925$	.357
Student Teaching Experience	$t(110) = -1.004$	.318	$U = 194.50, Z = -1.003$	.316

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Moderated Multiple Regression

The second research question asked: “What is the impact of assessment confidence on the relationship between pre-service teachers' assessment literacy and performance assessment scores?” The second research question sought to determine if there was a relationship between assessment content knowledge (i.e., Total CALI scores) and edTPA performance (i.e., edTPA Total scores and edTPA Assessment scores). Furthermore, the second research question asked if assessment confidence influenced this relationship between assessment knowledge and edTPA performance. In order to investigate this question, a series of six Moderated Multiple Regressions were analyzed. The regression results are presented below following the demographic and descriptive statistics, outlier and assumption tests, and bivariate correlations (i.e., between all variables in the models).

Table 43

Main Moderated Multiple Regression Models

Model Name	Independent Variable	Moderator	Dependent Variable
Model 1	CALI Total	Confidence Total	edTPA Total
Model 2	CALI Total	Confidence Total	edTPA Assessment
Model 3	Component 1 CALI	Component 1 Confidence	edTPA Total
Model 4	Component 1 CALI	Component 1 Confidence	edTPA Assessment
Model 5	Component 2 CALI	Component 2 Confidence	edTPA Total
Model 6	Component 2 CALI	Component 2 Confidence	edTPA Assessment

Model variables. Pertaining to the variables used in these analyses, there were two Dependent Variables (DV) examined in a series of separate models – the edTPA Total score and the edTPA Assessment score. The edTPA spans 15 rubrics segmented into three major domains (i.e., Planning, Instruction, and Assessment), which are called "Tasks." These rubrics consist of five levels of performance scored from one to five with higher scores indicating that the individual is an accomplished novice teacher. There are 25 points possible for each domain and a total of 75 points across the entire edTPA exam. Thus, the main DVs in this study included an edTPA Total score ranging from 15 to 75, and an edTPA Assessment score (i.e., the Assessment domain or task) ranging from five to 25.

For the main Independent Variables (IVs), the Total CALI (KR20 = .649), CALI Component 1 (KR-20 = .682), and CALI Component 2 (KR20 = .833) scores were examined in separate models. The Total CALI consists of 25 multiple-choice items scored either 0 (i.e., "Incorrect") or 1 (i.e., "Correct") with higher scores indicating more assessment knowledge or literacy. The CALI Components 1 and 2 contain 12 and 13 items, respectively, with the same multiple-choice items and scoring noted above. Therefore, these main IVs had potential score ranges from zero to 25 for the CALI Total, zero to 12 for CALI Component 1, and zero to 13 for CALI Component 2.

The moderator variables included the Confidence Total (Cronbach's α = .923), Component 1 Confidence (Cronbach's α = .889), and Component 2 Confidence scores (Cronbach's α = .833). The Confidence Total score consists of 25 items on a 5-point Likert scale paired with each CALI item. The confidence items are worded, "How confident are you in your response?" The response options on the Likert scale included, "Completely Unconfident" (Coded 0), "Mostly Unconfident" (Coded 1), "Neither Confident nor Unconfident" (Coded 2), "Mostly Confident" (Coded 3), and "Completely Confident" (Coded 4). Higher scores on these items are indicative or more assessment confidence related to assessment knowledge or literacy. Similar to the CALI Components, Confidence Components 1 and 2 contain 12 and 13 items, respectively, and the same Likert response scale detailed above. After averaging all total and components scores, these main moderators had potential ranges from zero to four (i.e., from the 5-point Likert confidence scale).

Table 44 (see below) summarizes the demographic and descriptive variables using the outcome variables in the Moderated Multiple Regressions. Across all the models, demographic variables were included as statistical controls (i.e., covariates). To support the use of these covariates in the models, the demographic and descriptive variable groups were reviewed and compared on edTPA Total and edTPA Assessment scores. Nonparametric tests were used if assumptions were violated such as nonnormality and unequal variances. Pearson or Spearman correlations were used to examine relationships between two continuous or ordinal variables, respectively. Independent *t*-Tests and One-Way ANOVAs (i.e., or nonparametric tests, if needed) were used to investigate edTPA Total and edTPA Assessment score differences between groups with two or three or more levels of the categorical variable. Any statistically significant relationships between the demographic and descriptive variables and the two edTPA outcomes were selected to be covariates in the Moderated Regression models.

The results indicated that there were significant relationships between Gender, Program, and GPA and the two main outcome variables in the model. For Gender, Females had higher average edTPA Total scores ($M = 42.60$, $SD = 7.86$) and higher average edTPA Assessment scores ($M = 13.81$, $SD = 3.93$; $t[95] = 3.71$, $p < .001$) compared to Males (Total: $M = 34.07$, $SD = 8.55$; Assessment: $M = 9.60$, $SD = 3.40$; $t[99] = 3.91$, $p < .001$). For Program, the omnibus tests indicated that there were statistically significant differences between the four program groups on edTPA Total scores ($F = 19.93$, $df = 3, 93$, $p < .001$) and edTPA Assessment scores ($F = 15.10$, $df = 3, 97$, $p < .001$). Post hoc comparisons on both edTPA outcome variables between programs revealed that AYA participants had significantly lower scores compared to all other groups ($p \leq .004$ for all). Additionally, there were significant differences on both edTPA scores between the ECED program and Other programs as well. The Other programs group in this study is comprised of all programs outside of ECED, MCED, and AYA and included Art Education, Music Education, Special Education, and Teaching English as a Second Language. Participants in the Other category had significantly lower edTPA Total scores and edTPA Assessment scores compared to the ECED program ($p \leq .001$ for all). Finally, GPA was significantly and positively related to both edTPA Total scores ($r = .381$, $p < .001$) and edTPA Assessment scores ($r = .395$, $p < .001$). Thus, two categorical variables (i.e., Gender and Program), and one continuous variable (i.e., GPA) were included as statistical controls in all six Moderated Multiple Regression models.

Table 44

Second Phase Sample Variable Descriptive Statistics: edTPA Overall Scores and edTPA Assessment Domain Scores (N = 112)

Variable	*n*	edTPA (Overall)			edTPA Assessment (Domain)		
		M/*Mdn*	*SD*/IQR	Min/Max	*M*/*Mdn*	*SD*/IQR	Min/Max
Gender							
Male	20	34.07	8.55	19/49	9.60	3.39	4/16
Female	92	42.60	7.86	25/62	13.81	3.93	5/21
Age	112	23.28/22.00*	2.77/2.00	20/35	--	--	--
Race							
White/Caucasian	102	41.38	8.59	19/62	13.16	4.18	4/21
Other	10	41.25	7.55	30/52	13.56	3.68	9/20
1st Generation College Student							
Yes	35	41.19	7.09	26/52	13.35	3.76	5/20
No	77	41.46	9.16	19/62	13.10	4.32	4/21
GPA	112	3.62/3.69*	.27/.42	2.96/4.00	--	--	--
Program							
ECED	34	46.76	5.67	38/62	15.78	3.07	5/21
MCED	14	43.38	9.26	26/61	14.31	4.29	6/21
AYA	29	32.43	6.79	19/50	9.57	3.69	4/19
Other	35	40.83	6.48	29/54	12.38	3.33	6/18
Year							
Senior	102	41.62	8.52	19/62	13.29	4.15	4/21
Other	10	37.67	7.45	28/48	11.50	3.56	6/15
Mother's Education							
HSD/GED or Less	27	42.67	7.46	29/52	13.74	4.10	5/20
Some College/Associate/Tech	29	40.23	8.61	25/62	13.13	3.79	6/20
Bachelor's Degree	40	41.41	9.23	19/61	13.01	4.29	4/21
Master's/Doctoral/Professional	16	40.80	8.64	26/52	12.58	4.73	6/20

Father's Education							
HSD/GED or Less	32	43.80	7.76	26/62	13.57	3.89	5/20
Some College/Associate/Tech	34	41.94	7.15	28/58	13.59	3.77	5/21
Bachelor's Degree	31	40.69	9.64	19/61	13.48	4.35	5/21
Master's/Doctoral/Professional	15	37.00	10.76	23/54	10.23	4.46	4/18
Course(s) with Assessment							
Yes Took Course(s)	101	41.92	8.44	19/62	13.37	4.08	5/21
No Did Not	11	37.09	7.79	23/52	11.73	4.36	4/20
Assessment-Specific Course							
Yes Took Course	30	42.19	7.89	30/61	13.21	3.72	6/21
No Did Not	82	41.06	8.72	19/62	13.18	4.29	4/21
Student Teaching Experience							
Yes	107	41.57	8.56	19/62	13.27	4.18	4/21
No	5	36.75	4.57	32/43	11.75	2.06	10/14
edTPA Overall	112	41.37	8.47	19/62	--	--	--
edTPA Assessment Domain	112	--	--	--	13.19	4.12	4/21

Note. Groups and continuous variables denoted with asterisks next to the values in the *M/Mdn* columns indicate non-normal distributions.

Table 45

Relationships between Second Phase Sample Variables and edTPA Overall Scores and edTPA Assessment Domain Scores (N = 112)

Variable	edTPA (Overall)		edTPA Assessment (Domain)	
	Statistical Test	p	Statistical Test	p
Gender	$t(110) = 3.708$	<.001***	$t(110) = 3.905$	<.001***
Age	$r = -.106$	.293	$r = -.175$	.081
Race	$t(110) = -.042$	.967	$t(110) = .266$	.790
1st Generation	$t(110) = .145$	.885	$t(110) = -.285$	.766
GPA	$r = .334$	<.001***	$r = .355$	<.001***
Program	$F(3,108) = 19.934$	<.001***	$F(3,108) = 15.096$	<.001***
Year	$t(110) = 1.107$	.271	$t(110) = 1.035$	.303
Mother's Education	$F(3,108) = .370$	.775	$F(3,108) = .265$	.851
Father's Education	$F(3,108) = 1.178$	.322	$F(3,108) = 2.203$	.093
Course(s) with Assessment	$t(110) = -1.800$	.075	$t(110) = -1.249$	.215
Assessment-Specific Course	$t(110) = -.586$	.560	$t(110) = -.039$	.969
Student Teaching Experience	$t(110) = -1.115$	.268	$t(110) = .710$	.479

Note. $^{*}p < .05$, $^{**}p <.01$, $^{***}p < .001$.

Outliers. The data were examined for outliers to enhance statistical conclusion validity. The residual diagnostics (i.e., studentized residuals) did not render any unusual outcomes (i.e., Y values in the regression formula) for cases. Large studentized residuals (i.e., > +3.0) indicate poor prediction of Y for each case, or extreme values for the outcomes with regard to the predictors in the equation. Cook's D was also consulted for extreme values (i.e., values close to 1 or 2 indicate potential problems). Next, Mahalanobis Distance values were examined; however, there were no cases that were strongly influential in the model. Finally, Leverages were calculated, and all values were less than .20, where values greater than .20 suggest influential data points (i.e., extreme values). After examining all the above outlier detection statistics, there were no outlying cases removed from the data.

Assumptions. Prior to conducting a Multiple Regression Analysis, several statistical assumptions were explored. The main statistical assumptions for Multiple Regression include: (1) Independence, (2) Normality, (3) Linearity, and (4) Homoscedasticity (Keith, 2006). The assumption of Independence states that the variance in the variable is independent, as opposed to the observed scores. The Durbin-Watson statistic can be used to identify if this assumption has been met (Hinkle, Wiersma, & Jurs, 2005). For both dependent variables, edTPA Total scores and edTPA Assessment scores, the assumption was met as the Durbin-Watson statistic fell between 1.5 and 2.5 (Durbin & Watson, 1951). The assumption of Normality tested the normal distribution of residuals in the data across the dependent variables (Shapiro & Wilk, 1965). The errors appeared to be normally distributed by viewing the histogram of the standardized residuals for both outcomes. Skewness statistics were also consulted (i.e., skewness divided by the standard error) to confirm the residual plot findings.

The Linearity assumption tests the linear relationship between the independent and dependent variables. Examining residual scatterplots of Y indicated that linearity was met. Finally, the Homoscedasticity assumption was assessed by examining patterns of data across the entire line of fit (Keith, 2006). The scatterplots of predictors and the DVs exhibited fairly constant dispersion of the values around the regression line for all values of X. Although not considered an assumption, multicollinearity was examined via a correlation matrix. Multicollinearity occurs when two or more independent variables are highly correlated (Keith, 2006). The tolerances for all the predictors were within acceptable limits with the variance inflation factors (VIFs) corroborating this evidence.

The collinearity diagnostics did not indicate any overlap in the contribution of the percentages to the model, and the condition indices were all within acceptable limits (i.e., < 30).

Correlations. Bivariate correlations were examined between all the main continuous variables in the model prior to the Multiple Moderated Regression analyses. Pearson correlations (*r*) were used for two continuous variables, and Spearman correlations (r_s) were conducted if a variable was significantly skewed. For the edTPA Total score models, CALI Total score was significantly and positively related to the hypothesized moderator Assessment Confidence Total score (r_s = .593, *p* < .001). CALI Total was also significantly and positively related to both edTPA Total score (*r* = .257, *p* < .011) and edTPA Assessment score (*r* = .267, *p* < .007). For the first component model, CALI Component 1 score was significantly and positively related to the hypothesized moderator Assessment Confidence Component 1 score (r_s = .599, *p* < .001). CALI Component 1 score was also significantly and positively related to edTPA Total score (r_s = .283, *p* < .005) and edTPA Assessment score (r_s = .280, *p* < .005). Finally, for the second component model, CALI Component 2 score was significantly and positively related to the hypothesized moderator Component 2 Confidence score (r_s = .423, *p* < .001). However, there were no significant relationships between this component and the two Dependent Variables – edTPA Total score (*p* = .211) and edTPA Assessment score (*p* = .121).

Table 46

Pearson and Spearman Correlation Matrix for Second Phase Outcome Variables (N = 112)

Variable	1	2	3	4	5	6	7	8	9
1. CALI Total	1.00								
2. CALI Component 1$_s$	.867***	1.00							
3. CALI Component 2	.796***	.470***	1.00						
4. Confidence Total$_s$	.593***	.543***	.462***	1.00					
5. Confidence Component 1$_s$	.619***	.599***	.443***	.945***	1.00				
6. Confidence Component 2$_s$	.480***	.403***	.423***	.921***	.757***	1.00			
7. edTPA Total	.257**	.282**	.128	.124	.140	.094	1.00		
8. edTPA Assessment	.267**	.282**	.155	.139	.167	.094	.936***	1.00	
9. Grade Point Average (GPA)	.275**	.304**	.140	.142	.136	.133	.381***	.395***	1.00

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$. Variable names with a subscript "s" were significantly skewed and Spearman correlations were reported.

Moderated Multiple Regression Models

In order to answer the second research question, a series of Moderated Multiple Regressions were conducted. There were six total regressions, and these were categorized into three groups based on the predictor in the model. These three groups contain two regressions each with the following labels: (1) Overall Assessment Knowledge Models (i.e., Models 1 and 2), (2) First Component Assessment Knowledge Models (i.e., Models 3 and 4), and (3) Second Component Assessment Knowledge Models (i.e., Models 5 and 6). In the first group (i.e., Overall), both regressions had the same focal predictor and moderator – CALI Total scores and Assessment Confidence scores. However, for Model 1, the outcome was edTPA Total scores and for Model 2, the outcome variable was edTPA Assessment scores. In the second group of regressions (i.e., First Component), the same main predictor (i.e., CALI Component 1 scores) and moderator (i.e., Component 1 Confidence scores) were included. The DVs for these regression models differed as described in the previous group above. That is, for Model 3 and Model 4, the outcomes were edTPA Total scores and edTPA Assessment scores, respectively. Finally, in the third group of regressions (i.e., Second Component), CALI Component 2 scores (i.e., main IV) and Component 2 Confidence scores (i.e., moderator) were included in both models. However, Model 5 (i.e., edTPA Total scores) and Model 6 (i.e., edTPA Assessment scores) in this group of regressions had different DVs.

A total of six Moderated Multiple Regressions were run with differing combinations of IVs (i.e., CALI Total, Component 1, or Component 2), Moderators (i.e., Assessment Confidence Total, Confidence Component 1, or Confidence Component 2) and DVs (i.e., edTPA Total or edTPA Assessment). Consistent across all models was the inclusion of three covariates – Gender (i.e., Females [Coded 1] and Males [Coded 0]), Program (i.e., Indicator Dummy Coding with ECED as the Reference Group [Coded 0 for All]), and GPA (i.e., a continuous variable ranging from 0 to 4.0). The hypothesis that confidence (i.e., Total or Component-specific) moderates the relationship between assessment literacy and performance assessment was examined. In addition, mean centering of variables and products were conducted, and the unstandardized coefficients were rendered and reported (Hayes, 2013). An a priori power analysis was conducted using G*Power 3 (Power = .80, α = .05; Faul, Erdfelder, Lang, & Buchner, 2007). The software tool yielded a minimum total sample size of 77 to detect medium effects (Cohen, 1988) in each model. Finally, each Moderated Multiple Regression analysis was conducted using Model 1 (with Covariates) in PROCESS Version 3 for SPSS developed by Hayes (2012).

Overall assessment knowledge models – model 1. As part of Research Question 2, Model 1 addressed if the relationship between CALI Total scores (i.e., overall assessment knowledge) and edTPA Total scores (i.e., overall assessment performance) changes depending on Assessment Confidence Total scores (i.e., overall assessment confidence) controlling for gender, program, and GPA. The predictors (i.e., main IVs and covariates), including the interaction, were tested using 95% bias corrected bootstrap confidence intervals (95% BCCI) and 1,000 bootstrap samples. Model 1 of Overall Assessment Knowledge was statistically significant (R^2 = .517, *F*[8, 88] = 12.103,

$p < .001$), with 51.8% of the variance in assessment performance explained by the predictors. In the full model, only the covariates of GPA (B = 8.069, SE = 3.953; t[88] = 2.041, p = .044, [95% CI: .214 to 15.926]), AYA Program status (B = -12.801, SE = 2.04; t[88] = -6.264, p < .001, [95% CI: -16.862 to -8.740]), and Other Program status (B = - 4.019, SE = 1.621; t[88] = -2.479, p = .015, [95% CI: -7.240 to -.798]) were significant predictors of edTPA Total scores. Assessment Confidence (i.e., interaction between edTPA Total scores and Assessment Confidence) was not a significant moderator of the relationship between overall assessment knowledge and performance assessment (ΔR^2 = .013, F[1, 88] = .974, p = .326, [95% CI: -.504 to 1.497]). That is, the relationship between pre-service teachers' assessment knowledge as measured by the modified CALI and performance as measured by the edTPA total score does not change depending on their level of assessment confidence, controlling for Gender, Program, and GPA (see Table 47).

Overall assessment knowledge models – model 2. For Research Question 2, Model 2 addressed if the relationship between CALI Total scores (i.e., overall assessment knowledge) and edTPA Assessment scores (i.e., domain-specific assessment performance) changes depending on Assessment Confidence Total scores (i.e., overall assessment confidence) controlling for gender, program, and GPA. The predictors (i.e., main IVs and covariates), including the interaction, were tested using the same bootstrap confidence interval specifications as in the first model (95% BCCI). Model 2 of Overall Assessment Knowledge was statistically significant (R^2 = .454, F[8, 92] = 11.855, p < .001), with 45.4% of the variance in assessment performance explained by the predictors.

In the full model, only the covariates of AYA Program status (*B* = -5.412, *SE* = 1.057; *t*[88] = -5.116, *p* < .001, [95% CI: -7.513 to -3.311]), and Other Program status (*B* = -2.355, *SE* = .811; *t*[88] = -2.904, *p* = .046, [95% CI: -3.966 to -.745]) were significant predictors of edTPA Assessment scores. Assessment Confidence (i.e., interaction between edTPA Assessment scores and Assessment Confidence) was not a significant moderator of the relationship between overall assessment knowledge and performance assessment (ΔR^2 = .008, *F*[1, 92] = 1.257, *p* = .326, [95% CI: -.1531 to .550]). That is, the relationship between pre-service teachers' assessment knowledge as measured by the modified CALI and performance as measured by the edTPA Assessment score does not change depending on their level of assessment confidence, controlling for Gender, Program, and GPA (see Table 47).

First component assessment knowledge models – model 3. For Research Question 2, Model 3 addressed if the relationship between Component 1 CALI scores (i.e., assessment knowledge) and edTPA Total scores (i.e., domain-specific assessment performance) changes depending on Component 1 Confidence scores (i.e., assessment confidence) controlling for gender, program, and GPA. The predictors (i.e., main IVs and covariates), including the interaction, were tested using the same bootstrap confidence interval specifications as in the previous models (95% BCCI). Model 3 of Component 1 Assessment Knowledge was statistically significant (R^2 = .524, *F*[8, 88] = 12.891, *p* < .001), with 52.4% of the variance in assessment performance explained by the predictors. In the full model, only the covariates of AYA Program status (*B* = -12.736, *SE* = 2.065; *t*[88] = -6.166, *p* < .001, [95% CI: -16.841 to -8.632]), Other Program status (*B* = -4.072,

$SE = 1.596$; $t[88] = -2.551$, $p = .013$, [95% CI: -7.245 to -.899]), and GPA ($B = 8.231$, $SE = 3.756$; $t[88] = 2.191$, $p = .031$, [95% CI: .766 to 15.695]) were significant predictors of edTPA Total scores. Component 1 Confidence (i.e., interaction between CALI Component 1 scores and Component 1 Confidence scores) was not a significant moderator of the relationship between overall assessment knowledge and performance assessment ($\Delta R^2 = .0172$, $F[1, 88] = 1.108$, $p = .295$, [95% CI: -.685 to 2.230]). That is, the relationship between pre-service teachers' assessment knowledge as measured by Component 1 CALI scores and performance as measured by edTPA total scores does not change depending on their level of assessment confidence, controlling for Gender, Program, and GPA (see Table 48).

First component assessment knowledge models – model 4. For Research Question 2, Model 4 addressed if the relationship between Component 1 CALI scores (i.e., assessment knowledge) and edTPA Assessment scores (i.e., domain-specific assessment performance) changes depending on Component 1 Confidence scores (i.e., assessment confidence) controlling for gender, program, and GPA. The predictors (i.e., main IVs and covariates), including the interaction, were tested using the same bootstrap confidence interval specifications as in the previous models (95% BCCI). Model 4 of Component 1 Assessment Knowledge was statistically significant ($R^2 = .462$, $F[8, 92] = 13.456$, $p < .001$), with 46.2% of the variance in assessment performance explained by the predictors. In the full model, only the covariates of AYA Program status ($B = -5.425$, $SE = 1.038$; $t[92] = -5.229$, $p < .001$, [95% CI: -7.491 to -3.367]), Other Program status ($B = -2.390$, $SE = .814$; $t[92] = -2.933$, $p = .004$, [95% CI: -4.008 to -.772]), and GPA (B

= 4.114, *SE* = 1.924; *t*[92] = 2.136, *p* = .035, [95% CI: .290 to 7.932]) were significant predictors of edTPA Assessment scores. Component 1 Confidence (i.e., interaction between CALI Component 1 scores and Component 1 Confidence scores) was not a significant moderator of the relationship between overall assessment knowledge and performance assessment (ΔR^2 = .0148, *F*[1, 92] = 2.243, *p* = .137, [95% CI: -.115 to .821]). That is, the relationship between pre-service teachers' assessment knowledge as measured by Component 1 CALI scores and performance as measured by edTPA Assessment scores does not change depending on their level of assessment confidence, controlling for Gender, Program, and GPA (see Table 48).

Second component assessment knowledge models – model 5. For Research Question 2, Model 5 addressed if the relationship between Component 2 CALI scores (i.e., assessment knowledge) and edTPA Total scores (i.e., domain-specific assessment performance) changes depending on Component 2 Confidence scores (i.e., assessment confidence) controlling for gender, program, and GPA. The predictors (i.e., main IVs and covariates), including the interaction, were tested using the same bootstrap confidence interval specifications as in the previous models (95% BCCI). Model 5 of Component 2 Assessment Knowledge was statistically significant (R^2 = .500, *F*[8, 88] = 12.345, *p* < .001), with 50.0% of the variance in assessment performance explained by the predictors. In the full model, only the covariates of AYA Program status (*B* = -12.796, *SE* = 1.932; *t*[88] = -6.621, *p* < .001, [95% CI: -16.637 to -8.956]), Other Program status (*B* = -4.065, *SE* = 1.612; *t*[88] = -2.511, *p* = .014, [95% CI: -7.283 to -.847]), and GPA (*B* = 8.059, *SE* = 3.954; *t*[88] = 2.038, *p* = .045, [95% CI: .202 to 15.918]) were significant predictors of edTPA Total scores. Component 2 Confidence (i.e., interaction between CALI Component 2 scores and Component 2 Confidence scores) was not a significant moderator of the relationship between overall assessment knowledge and performance assessment (ΔR^2 = .001, F[1, 88] = .143, p = .706, [95% CI: -1.001 to 1.473]). That is, the relationship between pre-service teachers' assessment knowledge as measured by Component 2 CALI scores and performance as measured by edTPA total scores does not change depending on their level of assessment confidence, controlling for Gender, Program, and GPA (see Table 49).

Second component assessment knowledge models – model 6. For Research Question 2, Model 6 addressed if the relationship between Component 2 CALI scores (i.e., assessment knowledge) and edTPA Assessment scores (i.e., domain-specific assessment performance) changes depending on Component 2 Confidence scores (i.e., assessment confidence) controlling for gender, program, and GPA. The predictors (i.e., main IVs and covariates), including the interaction, were tested using the same bootstrap confidence interval specifications as in the previous models (95% BCCI). Model 6 of Component 2 Assessment Knowledge was statistically significant (R^2 = .665, F[8, 92] = 12.520, p < .001), with 66.5% of the variance in assessment performance explained by the predictors. In the full model, only the covariates of AYA Program status (B = -5.425, SE = 1.048; t[92] = -5.172, p < .001, [95% CI: -7.507 to -3.341]), and Other Program status (B = -2.474, SE = 1.048; t[92] = -3.199, p = .002, [95% CI: -4.010 to -.938]) were significant predictors of edTPA Assessment scores. Component 2 Confidence (i.e., interaction between CALI Component 2 scores and Component 2 Confidence scores) was not a significant moderator of the relationship between overall assessment knowledge and performance assessment (ΔR^2 = .0001, F[1, 92] = .039, p = .847, [95% CI: -.531 to .436]). That is, the relationship between pre-service teachers' assessment knowledge as measured by Component 2 CALI scores and performance as measured by edTPA Assessment scores does not change depending on their level of assessment confidence, controlling for Gender, Program, and GPA (see Table 49).

Table 47

Moderated Multiple Regressions Models 1 and 2: Predicting edTPA Total and Assessment Scores from CALI Total Scores Moderated by Assessment Confidence Scores (N = 112)

Model and Variables	*B*	*SE*	*t*	LLCI	ULCI	R^2	*F*
Model 1 (*N* = 97)							
Constant	15.927	14.65	1.087	-13.187	45.040		
Confidence Score	1.900	1.428	1.331	-.938	4.739		
CALI Score	.139	.269	.517	-.396	.674	.518	12.103^{***}
CALI X Confidence	.497	.503	.987	-.504	1.498	(Δ) .013	.974
Gender							
Female (0)	--	--	--	--	--		
Male (1)	-1.829	2.242	-.816	-6.284	2.626		
Program							
ECED (0)	--	--	--	--	--		
MCED (1)	-1.473	2.489	-.592	-6.420	3.473		
AYA (1)	-12.801	2.044	-6.264^{***}	-16.862	-8.740		
Other (1)	-4.019	1.621	-2.479^{*}	-7.240	-.798		
GPA	8.070	3.953	2.041^{*}	.214	15.926		
Model 2 (*N* = 101)							
Constant	.927	7.422	.124	-13.875	15.669		
Confidence Score	.641	.701	.913	-.752	2.034		
CALI Score	.097	.148	.657	-.197	.395	.454	11.855^{***}
CALI X Confidence	.198	.177	1.12	-.153	.550	(Δ) .008	1.257
Gender							
Female (0)	--	--	--	--	--		
Male (1)	-1.1897	1.10	-1.07	-3.39	1.01		
Program							
ECED (0)	--	--	--	--	--		
MCED (1)	-.565	1.18	-.477	-2.92	1.788		
AYA (1)	-5.41	1.05	-5.11^{***}	-7.51	-3.311		
Other (1)	-2.35	.811	-2.90^{*}	-3.96	-.744		

GPA	3.90	2.01	1.94	-.089	7.895		

Note. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. *B* = Unstandardized regression coefficient.

Table 48

Moderated Multiple Regressions Models 3 and 4: Predicting edTPA Total and Assessment Scores from Component 1 CALI Scores Moderated by Component 1 Assessment Confidence Scores (N = 112)

Model and Variables	*B*	*SE*	*t*	LLCI	ULCI	R^2	*F*
Model 3 (*N* = 97)							
Constant	15.161	14.054	1.078	-12.768	43.091		
Confidence Score	1.68	1.200	1.401	-.703	4.067		
CALI Score	.316	.3810	.829	-.441	1.073	.523	12.890***
CALI X Confidence	.772	.733	1.053	-.6853	2.230	(Δ) .017	1.108
Gender							
Female (0)	--	--	--	--	--		
Male (1)	-1.674	2.305	-.726	-6.256	2.907		
Program							
ECED (0)	--	--	--	--	--		
MCED (1)	-1.688	2.466	-.684	-6.589	3.213		
AYA (1)	-12.736	2.065	-6.166***	-16.841	-8.632		
Other (1)	-4.072	1.596	-2.551**	-7.245	-.899		
GPA	8.231	3.756	2.191*	.766	15.695		
Model 4 (*N* = 101)							
Constant	.047	7.146	.006	-14.146	14.241		
Confidence Score	.898	.583	1.538	-.261	2.057		
CALI Score	.112	.190	.587	-.266	.490	.461	13.456***
CALI X Confidence	.353	.235	1.497	-.115	.821	(Δ) .014	2.243
Gender							
Female (0)	--	--	--	--	--		
Male (1)	-1.117	1.10	-1.011	-3.310	1.076		
Program							

ECED (0)	--	--	--	--	--
MCED (1)	-.596	1.198	-.497	-2.976	1.784
AYA (1)	-5.429	1.038	-5.229***	-7.491	-3.367
Other (1)	-2.390	.814	-2.933**	-4.008	-.7718
GPA	4.111	1.924	2.136*	.290	7.932

Note. *p < .05; **p < .01; ***p < .001. *B* = Unstandardized regression coefficient.

Table 49

Moderated Multiple Regressions Models 5 and 6: Predicting edTPA Total and Assessment Scores from Component 2 CALI Scores Moderated by Component 2 Assessment Confidence Scores (N = 112)

Model and Variables	*B*	*SE*	*t*	LLCI	ULCI	R^2	*F*
Model 5 (*N* = 97)							
Constant	16.582	14.677	1.129	-12.585	45.750		
Confidence Score	.910	1.376	.661	-1.823	3.645		
CALI Score	.128	.497	.257	-.861	1.117	.500	12.345***
CALI X Confidence	.235	.622	.378	-1.001	1.473	(Δ) .001	.143
Gender							
Female (0)	--	--	--	--	--		
Male (1)	-2.584	2.408	-1.073	-7.370	2.201		
Program							
ECED (0)	--	--	--	--	--		
MCED (1)	-1.387	2.639	-.525	-6.631	3.857		
AYA (1)	-12.796	1.932	-6.621***	-16.637	-8.956		
Other (1)	-4.065	1.619	-2.510*	-7.283	-.847		
GPA	8.059	3.954	2.038*	.201	15.917		
Model 6 (*N* = 101)							
Constant	1.443	7.412	.194	-13.277	16.165		
Confidence Score	.011	.529	.021	-1.039	1.062		
CALI Score	.204	.262	.777	-.317	.725	.442	12.520***
CALI X Confidence	-.047	.243	-.196	-.531	.435	(Δ) .001	.038

Gender					
Female (0)	--	--	--	--	--
Male (1)	-1.663	1.163	-1.430	-3.973	.646
Program					
ECED (0)	--	--	--	--	--
MCED (1)	-.641	1.254	-.511	-3.133	1.850
AYA (1)	-5.424	1.048	-5.172***	-7.507	-3.341
Other (1)	-2.474	.773	-3.199**	-4.010	-.938
GPA	3.869	1.989	1.944	-.081	7.820

Note. $^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$. B = Unstandardized regression coefficient

Summary

The aims of Chapter 4 were to investigate the psychometric properties (i.e., reliability and validity) of the modified CALI, evaluate the underlying structure of the measure, and explore the relationship between assessment confidence, knowledge, and performance. While the modified CALI appeared to be psychometrically acceptable, the dimensionality of the measure was inconclusive. Analyses supported both a 25-item unidimensional measure, as well as an underlying a two-component structure. Additionally, the moderating relationship of confidence between knowledge and performance was not found. However, other variables such as program and GPA were found to have significant impacts on both assessment knowledge and performance. The following chapter (Chapter 5: Discussion) discusses these findings and provides implications for teacher education programs, researchers, and statistical and psychometric proceedings. This subsequent chapter also presents the limitations and possibilities for future research.

CHAPTER V

DISCUSSION

The objectives of this study were twofold: (1) To examine the psychometric properties of an assessment literacy measure (i.e., the modified Classroom Assessment Literacy Inventory [CALI]; Mertler, 2003) and an assessment confidence measure (i.e., the modification/addition to the CALI) in an undergraduate teacher education program student sample (i.e., pre-service teachers), and (2) To investigate the relationship between assessment literacy, assessment confidence, and performance assessment (i.e., the edTPA portfolio-based assessment). More specifically, the psychometric properties (i.e., content and construct validity, internal consistency reliability) of the CALI and the newly-developed assessment confidence measure were examined to provide evidence of the internal structure (i.e., unidimensional or multidimensional) and score reliabilities. Rasch Analysis, Rasch Principal Components Analysis (PCA), and Confirmatory Factor Analysis (CFA) were used to analyze the assessment literacy and confidence items prior to including these measures as predictors (i.e., the main Independent Variable [IV] and Moderator) of high-stakes performance assessment scores.

The two study objectives outlined above correspond to the two main research questions in this study. The first study objective included both Research Question 1 and 1A, and the second study objective was related to Research Question 2. Research Question 1 (RQ1) stated, "What are the psychometric properties of the newly-developed assessment literacy and confidence measure for pre-service teachers?" This first research question contained an intentionally broad measurement term – "psychometric properties"

– that is typically used in exploratory measure development research in reference to reliability and validity. A related, more specific research question was included with RQ1. Research Question 1A stated, “What is the internal structure (i.e., unidimensional or multidimensional) of the modified CALI?” These broad (RQ1) and specific (RQ1A) research questions reflected the status of the modified CALI as containing both a newer, more exploratory component (i.e., assessment confidence) and an older, more confirmatory portion (i.e., assessment literacy). That is, the assessment literacy items had a preexisting foundation of theoretical and some applied, research-based support, as they were developed using the seven areas of teacher assessment knowledge from the *Standards*. However, the addition of confidence ratings for each assessment literacy item necessitated a comprehensive psychometric investigation of both constructs as the existing reliability and validity evidence is not applicable to the modified measure.

Research Question 2 (RQ2) stated, “What is the impact of assessment confidence on the relationship between pre-service teachers' assessment literacy and performance assessment scores?” This second research question investigated the influence of assessment confidence on the relationship between assessment knowledge and assessment performance (i.e., the edTPA portfolio-based assessment) using a series of Moderated Multiple Regressions. RQ2 extends the psychometric evidence from RQ1 and RQ1A by modeling the construct of confidence as a moderator that influences the relationship between literacy and performance. The results from this second research question contribute to teacher education preparation related to the edTPA performance-based assessment. In addition, the results also contribute theoretically to measurement research and practice by providing validity evidence for the scores on the confidence and literacy measures.

This chapter (Chapter 5) begins with a discussion of the results from the two main research questions (i.e., Research Questions 1 and 1A [RQ1 and RQ1A], and Research Question 2 [RQ2]). Expanding the discussion of RQ1 and 1A and RQ2, the implications from this study's results for use within the current settings are detailed, followed by the limitations and future research directions. Chapter 5 presents the evidence and conclusions in five sections: (1) Research Question 1 and 1A, (2) Research Question 2, (3) Implications, (4) Limitations and Future Directions – Conceptual, Methodological, and Statistical and Psychometric, and (5) Conclusion.

Research Question 1 and 1A

The first objective of this study was to investigate the psychometric properties (i.e., reliability and validity) of the modified CALI, specifically the item-level information for all multiple-choice content questions. The goal of the first research question was to evaluate the modified CALI as a measure of assessment literacy using a small sample of pre-service teachers. RQ 1 asked: "What are the psychometric properties of the newly-developed assessment literacy and confidence measure for pre-service teachers?" A subsequent, more specific research question (1A) posited, "What is the internal structure of the modified CALI?" The following sections contain a summary description of the pilot phase sample demographic and descriptive information. Subsequent sections include comparisons between different unidimensional and multidimensional assessment literacy and assessment confidence measures.

Demographic and Descriptive Information (Pilot Sample)

The descriptive information in this section can be summarized into two major categories: Demographic and Academic. Demographic information includes gender, age, race, and mother's and father's highest level of education. The academic information includes year in school, first-generation college student status, teacher education program, cumulative GPA, students' enrollment in courses with an assessment component, students' enrollment in assessment-only courses, and if the pre-service teacher has had student teaching experience. The information for the pilot sample (*N* = 165) in this section will be summarized using these two major categories as will the information for the second phase sample for ease of comparison. For the demographic information, the overwhelming majority of the pilot sample was female and White/Caucasian, with an average age of 21. For highest level of education attained by students' parents, the sample's mothers and fathers had slightly more Bachelor's degrees and High School Degrees (HSD)/General Equivalency Diplomas (GED).

For the academic information, pilot sample students were predominantly Juniors and Seniors who were not first-generation college students, with a larger proportion in the ECED program and an average cumulative GPA of 3.5. A considerable proportion (i.e., over 70%) of students had enrolled in courses with an assessment component, with less than 20% indicating enrollment in assessment-only courses. Finally, an item on the modified CALI asked students to report their experience(s) in the classroom (i.e., "In your undergraduate program, did/do you have experience in the classroom in any of the following capacities? [Select all that apply]"). This sample contained more students responding “No” when asked if they have had any student teaching experience, although the groups performed nearly equivalent.

Participant demographic and academic information was examined in relation to the total modified CALI scores. The average total CALI score (out of 35 possible points) was just under 18.5. GPA was positively related to total modified CALI scores, and for race, White/Caucasian students (*n* = 157; *M* = 18.50, *SD* = 3.64; *Mdn* = 19.00, IQR = 4.00) had higher total CALI scores compared to others (*n* = 8, *M* = 15.63, *SD* = 4.34; *p* < .05). Interestingly, there were no differences on total assessment literacy scores between students with and without student teaching experience. This finding suggests that students with no student teaching experience were not at a disadvantage with regards to their level of assessment knowledge. Additionally, participants were asked if they had taken any courses containing assessment content (i.e., either as an entire course or as one component of a course) in an effort to investigate the impact of exposure to assessment content on assessment knowledge and confidence. However, the inferential tests conducted on two CALI items reported in Table 3 (i.e., “Have you ever taken a course in which the topic was only assessment?” and “Have you ever taken a course in which assessment was one of multiple topics covered?”) showed that taking an assessment-specific course or courses that contain a small assessment component did not have an impact on students’ assessment literacy scores (*p* > .05).

Descriptive information for assessment confidence within the pilot sample must also be discussed. Participant demographic and academic information was examined in relation to the total average assessment confidence scores. The total assessment confidence (on a scale of zero to four points as coded in the Likert-scale) was 2.55. For program, ECED and AYA students had higher average assessment confidence scores compared to MCED and students in all Other programs. When asked “Have you ever taken a course in which assessment was one of multiple topics covered?”, there was a difference on average assessment confidence scores between students who responded “Yes” and those who responded “No” (p = .003). These findings indicate that taking an assessment-related course did have an impact on assessment confidence in this sample. There were also differences in confidence according to gender, age, and mother’s education. Specifically, males (n = 45; M = 18.67, SD = 3.33) scored higher than females (n = 120; M = 18.24, SD = 3.85; Mdn = 19.00, IQR = 5.00; p < .05). Older students reported higher confidence and participants whose mothers had graduate-level degrees indicated that they had lower confidence (p < .05 for all). Additionally, students who were not first-generation college students felt significantly less confident than students who were first generation college students.

Assessment Literacy Measure Development (Pilot Sample)

The following sections will contain a summary and discussion of the pilot sample Rasch Analysis results. The results were organized in chronological order of analysis, beginning with the original 35-item modified CALI and assessment confidence results. The CALI and confidence analyses that followed the original 35-item analyses included the 25-item versions, and the two-component versions. For this discussion, the results will be organized by internal component structure – unidimensional and multidimensional. This reorganization corresponds to an important psychometric finding reoccurring throughout the study – there is support for both a unidimensional and two-component internal structure of the CALI. Thus, sections are organized in the following order: (1) Unidimensional Assessment Literacy: Comparing the 35-Item and 25-Item Measures, (2) Unidimensional Assessment Confidence: Comparing the 35-Item and 25-Item Measures, (3) Assessment Literacy Dimensionality: Rasch Principal Components Analysis (PCA), (4) Multidimensional Assessment Literacy: Comparing the Two-Component and 25-Item Unidimensional Measures, and (5) Multidimensional Assessment Confidence: Comparing the Two-Component and 25-Item Unidimensional Measures.

For Sections 1, 2, 4, and 5, the *Journal of Applied Measurement* "Guidelines for Manuscripts" (Smith, Linacre, & Smith, 2003) involving applications of Rasch measurement will be used to compare the differing iterations of the modified CALI and evaluate the collection of evidence for discussion. These guidelines include a comparison of the following for the Assessment Literacy items: (1) Item and Person Summary and Fit Statistics (Real), and (2) Item and Person Separation and Reliability (Real). The above guidelines are similar for Assessment Confidence with the addition of one section: (1) Rating Scale Functioning, (2) Item and Person Summary and Fit Statistics (Real), and (3) Item and Person Separation and Reliability (Real).

Unidimensional assessment literacy: comparing the 35-Item and 25-Item measures. The pilot Rasch analysis revealed overall low assessment literacy scores on the modified CALI. The average score was only 18.30 points out of a possible 35, or 52.3% correct responses. There were four items that no participant answered correctly and three additional difficult questions that only one person answered correctly. This indicates that the pilot version, the original 35-item CALI developed by Mertler (2005), contained difficult content questions for this sample. Overall, the reliability of the scores on the measure in this pilot phase of the study was high, which indicates the ability of the modified CALI items to appropriately place participants across all programs (i.e., ECED, MCED, AYA, and Other programs) along the continuum of assessment knowledge. While these psychometric qualities (i.e., reliability and validity) of the pilot were strong, the content was not at the appropriate level or relevance for this sample.

Ten poorly fitting items were removed at this stage and items were renumbered from one to 25. Then, the pilot sample Rasch Analysis of the 25-item modified CALI was conducted. This analysis revealed a slightly increased average score from the 35-item CALI with 59.2% correctly answered items (or 14.8 correct responses out of a possible 25). Items still were higher in difficulty relative to this sample's assessment knowledge. Overall, the reliability of the measure was high and at the same magnitude compared to the 35-item CALI. Additionally, there were only two item fit issues for consideration. Compared to the 35-item CALI, the psychometric properties (i.e., reliability and validity) of the 25-item measure were comparably strong. However, the content was slightly less difficult or better aligned with this sample on the 25-item CALI, and there was an increase in the average scores over the 35-item version.

Unidimensional assessment confidence: comparing the 35-Item and 25-Item measures. An additional pilot Rasch analysis of confidence data was also reported. This information provides the study with an understanding of the level of assessment confidence in this sample. In Rasch terms for the 35-item measure, the person mean was .64, with a logit range (i.e., value on the standardized ordinal continuum or item-person map) of -1.95 to 3.13, which indicates that overall this sample was somewhat confident, with minimal variance (Sample SD = 17.7; SE = 0.02), in their assessment knowledge. The Likert category structure of the confidence questions was also investigated. Consistent with the reported mean confidence level, 30% of all endorsements were for the neutral option “Neither Confident Nor Unconfident” and “Mostly Unconfident” comprised only 10% of responses. The majority of pre-service teachers, however, were above average with 68% of all confidence responses either neutral or mostly confident. Since this sample was overall confident, it is difficult to truly gauge the level of confidence given their tendency to endorse the same categories.

An additional analysis of confidence for the 25-item modified CALI was also reported. In Rasch terms, the pre-service teacher mean was .79, with a logit range of -1.66 to 3.00, which indicates that overall, this sample was even more confident in their assessment knowledge compared to using the 35-item measure. The category structure of the confidence questions showed that the sample had average confidence (i.e., 27% were for the neutral option “Neither Confident Nor Unconfident”) or higher (i.e., 40% were “Mostly Confident” and 21% were “Completely Confident”). This means that 67% of all confidence responses were either neutral or mostly confident. These results are consistent, but slightly higher than the response patterns found in the 35-item measure.

Comparing confidence differences between the 35-item and 25-item versions showed that pre-service teacher confidence increased from .64 to .79, and no new psychometric concerns were introduced. This substantial increase in average confidence indicates that by removing the difficult items, participants were more confident in the knowledge relevant to what they had learned. The extremely difficult items were those in which participants felt the least confidence in their answers, and therefore warranted removal (e.g., “The following standardized test data are reported for John. Subject Stanine Score: Vocabulary – 7, Mathematics Computation – 7, Social Studies – 7. Which of the following is a valid interpretation of a score report?”). The confidence response categories performed consistently with that of the 35-item version previously discussed, which indicates that the sample primarily endorsed two to five categories.

Assessment literacy dimensionality: Rasch Principal Components Analysis (PCA). The PCA analysis reported two components from the 25 items. Component 1 is comprised of 12 items and Component 2 is comprised of 13 items. Broad analyses of these components led to the identification of some basic common themes as well as psychometric qualities like the type of question (i.e., vignette or no vignette). With regards to content, the 12 items in Component 1 share the themes of scoring student performance and testing or assessment ethics. All items require the content knowledge for either determining how to grade student performance or understanding testing ethics at the classroom and standardized or state level. The 13 items in Component 2 address choosing assessment strategies and communicating assessment results. This domain includes choosing what type of assessments, items, and exercises to give students, as well as communicating and understanding basic measurement principles (e.g., percentiles, norming, reliability, ranking).

Basic descriptive analyses indicated that on average students performed better on the 12-item Component 1 reporting 68.0% correct responses compared to the 13-item Component 2 with only 49.0% correct responses. It can therefore be stated that within this sample, students had higher scores on items related to scoring student performance and testing or assessment ethics versus choosing assessment strategies and communicating assessment results. In comparing pre-service and in-service teachers, Mertler (2004) found that pre-service teachers performed consistently lower on CALI items connected to *Standards* involving scoring and ethics; however, the lowest scores regardless of teacher experience (i.e., pre-service vs. in-service) involved *Standards* related to selecting assessment strategies and communicating results. Therefore, previous research shows some support for the two-component structure in the current study, with broader categorization of the *Standards* across two assessment knowledge-related areas defined by assessment content knowledge and the application of knowledge.

Multidimensional assessment literacy: comparing the two-component and 25-item unidimensional measures. For Component 1, participants responded correctly to 71.7% of items on average, and for Component 2, the average was 47.7%. This finding indicates that Component 1 questions were easier for participants to correctly answer, and the average score on this component was higher than the 25-item version. On the other hand, the Component 2 average scores were lower than the 25-item version. Item and Person fit indices did not indicate potential psychometric problems for either component. Item separation and reliability were acceptable and similar to the 25-item version;

however, person reliability for Component 2 was low. This is likely due to the difficulty of Component 2 questions as indicated by the low average score.

With Component 1 considered “easier” than Component 2, it is unsurprising that the 25-item CALI (i.e., the unidimensional model) produced psychometric properties (i.e., reliability and validity) that were a combination of the two. This finding also suggests that the subtle differences in content between the Components may be a reflection of assessment knowledge related to topics covered in coursework (e.g., scoring) compared to assessment knowledge related to application of coursework knowledge (e.g., selecting strategies, communicating results). Applied knowledge may be conceptually more challenging for students in this sample; however, it is also possible that the application of knowledge may be more difficult to measure in this sample. Mertler and Campbell (2005), in their construction and modification of the ALI to the CALI acknowledged this when they included more application items with the present-day CALI containing both items reflective of classroom assessment knowledge from coursework and items targeting application benefitting those with more hands-on experience. Thus, there is support for the unidimensional 25-item CALI and the two-component CALI, as both have good psychometric properties (i.e., reliability and validity).

However, use of either version may depend on the target population. For example, it may be more appropriate to use the unidimensional CALI when testing a large group containing both pre-service and in-service teachers (i.e., the fundamental basic assessment concepts), and either Component 1 or Component 2 when testing pre-service teachers or in-service teachers on more difficult applied knowledge, respectively. Component 1 may be the better choice for pre-service teachers during their coursework in a teacher preparation program. However, Component 2 could also be administered as students apply their knowledge in settings such as student teaching, or as a way to gauge preparedness for an actual classroom environment upon graduation. Administration of Component 2 and the scores produced after students complete their student teaching may better represent their applied assessment knowledge, as the current sample was (on average) in the process of student teaching and had not completed their final semester of student teaching at that time in this study

Multidimensional assessment confidence: comparing the two-component and 25-item unidimensional measures. Assessment confidence on Component 1 averaged 34.4 points out of a possible 48 or 71.7% (i.e., 12 items on a scale coded from 0 to 4). Component 2 averaged 31.8 points out of a possible 52 or 61.2% (i.e., 13 items on a scale coded from 0 to 4). Fit statistics, item reliability, and person reliability all reported acceptable values for both components and for the unidimensional 25-item confidence measure. In addition, the rating scale response patterns for the two components were consistent with the unidimensional model. The majority of confidence responses fell between the neutral option, "Mostly Confident," and "Completely Confident."

The slightly elevated confidence for Component 1 (71.7%) compared to Component 2 (61.2%), is unsurprisingly considering the performance on the first Component was better. The positive relationship between knowledge and self-efficacy and/or confidence (i.e., as knowledge increases, so does self-efficacy and/or confidence) has been noted in the literature (e.g., Bursal & Paznokas, 2006; Garbett, 2003; Watson, 2001; Wolters & Daugherty, 2007). Additionally, these results highlight the importance of Bandura's (1977) theory on the relationship between confidence and/or self-efficacy and performance. Theoretically, confidence in more coursework-related assessment knowledge or the easier fundamental assessment concepts (Component 1) should be higher for pre-service teachers compared to confidence related to the more difficult application of assessment knowledge (Component 2). Therefore, the use of either the unidimensional or two-component measures of assessment confidence may depend on the target population and the purposes for using the results, as was discussed in the section above.

From Bandura's (1977) theory applied to teachers, those who feel competent and confident in their assessment abilities are more likely to engage in the process of assessment (Black & Wiliam, 1998). Practice, exposure, and application of coursework knowledge may contribute to how confident a teacher is when using or applying assessment-related skills in a new context. Confidence related to Component 1 (i.e., confidence in assessment knowledge) may be more accurately measured for new students in a teacher preparation program who spend the majority of their first few years studying material from their coursework. The themes in this component may contain "easier" content for students at this level to master or perhaps illustrate assessment basics. However, confidence related to the applied themes in Component 2 involves conceptually more challenging content. Therefore, measuring applied knowledge over time throughout the later years of pre-service teacher preparation programs may assist in monitoring the impact of the hands-on opportunities provided throughout the curriculum.

Research Question 2

The second objective of this study was to investigate the relationship between assessment literacy, assessment confidence, and performance assessment. The goal of RQ2 was to evaluate the hypothesized role of assessment confidence in the relationship between assessment literacy and assessment performance in pre-service teachers. RQ2 asked: "Does confidence moderate the relationship between modified CALI scores (i.e., assessment literacy knowledge) and edTPA performance (i.e., portfolio-based assessment knowledge)?" Prior to discussing this research question, the second phase sample characteristics are discussed. Next, using the second phase sample item responses to the modified 25-item CALI, the results from the Confirmatory Factor Analysis (CFA) are summarized and evaluated.

Demographic and Descriptive Information (Second Phase Sample)

The descriptive information in this section is again summarized into two major categories: Demographic (i.e., gender, age, race, and parental education level) and Academic (i.e., year in school, first-generation college student status, program, cumulative GPA, enrollment in assessment-only courses, enrollment in courses with an assessment component, and student teaching experience). The information for the second phase sample ($N = 112$) in this section will be summarized using these two major categories as was done for the pilot sample for ease of comparison. For the demographic information, the overwhelming majority of the second phase sample was female and

White/Caucasian, with an average age of 23. For highest level of education attained by students' mothers and fathers, the sample's mothers predominately had either a High School Degree (HSD)/General Equivalency Diploma (GED) or a Master's, Doctoral, or Professional terminal degree. For fathers' education, the majority held a Master's, Doctoral, or Professional terminal degree.

For the academic information, second phase sample students were mainly Seniors who were not first-generation college students, with a larger proportion in the ECED program or in a program in the Other category (e.g., Special Education, Music, Language) and an average cumulative GPA of 3.62. A considerable proportion (i.e., over 90%) of students had enrolled in courses with an assessment component, with less than 27% indicating enrollment in assessment-only courses. Information relative to enrollment in these courses according to programs was not considered at this stage of the study. Finally, this sample contained nearly 96% of students responding "Yes" when asked if they had any student teaching experience, which was expected as all students in this sample were in their final semester of an undergraduate teacher preparation program. The remaining 4% (*n* = 5) of student who responded "No" likely represents persons who misinterpreted the question, as all students had some student teaching experience as required by University X's graduation requirements.

Participant demographic and academic information was examined in relation to the total modified CALI scores. The average total CALI score (out of 25 possible points) was 14.54. GPA was positively related to total modified CALI scores. This indicates that higher student performance in coursework is related to higher scores on the modified

CALI. Additionally, more confidence in assessment knowledge was related to higher CALI performance. There were no differences on assessment knowledge reported between student exposure to assessment via coursework (i.e., from the CALI item “Have you ever taken a course in which assessment was one of multiple topics covered?”) or student teaching experience, although for both of these variables nearly all students in the sample reported some exposure to assessment and teaching experience.”

Descriptive information for assessment confidence within the second phase sample must also be discussed. Participant demographic and academic information was examined in relation to the total average assessment confidence scores. The total assessment confidence (on a scale of zero to four points as coded in the Likert-scale) was 2.67. For program, ECED and AYA students had higher average assessment confidence scores compared to MCED and students in all Other programs. There were no differences on average assessment confidence scores between students with and without student teaching experience and between students with exposure to assessment either in a course with an assessment component (i.e., the CALI item “Have you ever taken a course in which assessment was one of multiple topics covered?”) or in a course that focused only on assessment (i.e., the CALI item “Have you ever taken a course in which the topic was only assessment?”). These findings indicate that exposure to assessment did not impact assessment confidence, but at the program level there were differences in levels of confidence. Additionally, students who were not first-generation college students felt less confident than students who were first generation college students.

Assessment Literacy Measure Internal Structure Confirmation (Second Phase Sample)

Based on the structure evidenced within the pilot testing phase, the results of the full-scale administration were analyzed using factor analytic techniques. The results of the analyses were used to evaluate fit with the preliminary component structure and to assess the underlying structure of the modified CALI. The intention of the factor analytic approach was to validate the PCA results. The points of discussion presented here are the statistical evidence supporting the presence of two components, and the strength of that evidence in considering either a unidimensional or multidimensional internal structure.

Internal structure of the CALI with Confirmatory Factor Analysis (CFA). A CFA was conducted in order to confirm the two-component structure reported in the Rasch PCA analysis. After conducting parceling procedures to create four parcels for each of the two components, it took three models to achieve acceptable model fit for two components. All parcels, and therefore items, significantly loaded on their hypothesized components from the Rasch PCA. However, two additional model modifications were made. To begin, the first model modification required adding an error covariance between two parcels on the same component (e.g., Component 2). Since these parcels loaded on the same component throughout these analyses and the component is representative of a general content domain (e.g., choosing assessment methods and strategies and communicating assessment results.) the addition of the covariance was supported. However, the second modification added an error covariance across parcels in different components – and therefore suggests a connection between some content areas in both components.

Upon a further content-focused investigation, Item 15 contained elements indicative of creating assessments (Component 2) and scoring (Component 1). Additionally, Item 2 focused on understanding assessment results, which is logically related to using assessment results. While the two-component structure has some validity evidence, there is a clear degree of overlap or connection between some of the content in both components, which implies a weakness in the two-component model. Problems involving the content of items relative to the internal structure may be a product of the changing assessment context (e.g., accountability systems). Gotch and French (2014) suggest that researchers consider the representativeness and relevance of the items in light of recent state and national assessment-related transformations. These present-day changes are not reflected in measures such as the CALI that were constructed based on the *Standards* and may be contributing to the overlap and less fine-grained distinction between assessment knowledge content areas (Xu & Brown, 2016).

A final point of discussion probes the strength of the two-component model. It should be noted that the Rasch PCA analyses reported the minimum eigenvalue of approximately two for a two-component model. This minimum value of two suggests that the second component carries at least the weight of two items. As a 25-item measure, a second underlying component with only the strength of two items is the minimal amount of evidence needed to determine the clear presence of a second component. This issue was echoed in the CFA when the cross loading between items was used for model modification. Clearly, much of the content in the components is overlapping. Reflecting on this finding, it should be mentioned that all three questions related to ethics were subsumed under the first component. This was not true of any other content-similar set of items. These ethics items may be contributing to the evidence of a two component model.

Moderating the Assessment Literacy and Performance Assessment Relationship

The second research objective of this study investigated the role of confidence in assessment knowledge as measured by the modified CALI and how this is related to the edTPA exam. This objective aimed to determine the relationship between confidence and assessment knowledge on the performance assessment outcomes as observed on the edTPA. The statistical approach for this objective was to conduct a Moderated Multiple Regression investigating the impact of confidence (i.e., the Moderator) on assessment knowledge's relationship to the outcome (i.e., the edTPA total score or assessment score). Moderation Analysis answers the question of when (i.e., when does confidence impact edTPA performance). In the following paragraphs, the results of the moderation analysis are discussed. While confidence was not a significant moderator between assessment knowledge and performance, covariates used in these models present important findings for teacher preparation.

Moderated Multiple Regression models 1 through 6. Results from six different moderation models indicated that confidence did not significantly moderate the relationship between modified CALI scores, edTPA total scores or assessment scores. This is likely because of the close temporal presentation of each item related to either the moderator (confidence) or the target construct (assessment knowledge). The idea of temporal precedence was investigated due to the lack of moderation and it was discovered that conceptual moderation models address this issue. Temporal precedence means that the moderator must precede that target in order to accurately assess the presence of a moderator (Kraemer, Kiernan, Essex, & Kupfer, 2008). In the case of the modified CALI, the confidence questions appear directly after the assessment knowledge questions (i.e., the construct target). This ordering is not conducive to investigating true moderation because the moderator appeared in close proximity to the target.

The moderation models were not significant, but the covariates did produce important effects for discussion. The inclusion of covariates in each of these models was consistent with gender, cumulative GPA, and program used. While none of the moderation models were significant, when controlling for these covariates there were significant differences in performance on edTPA total scores and edTPA assessment scores. GPA and program (i.e., AYA and the “Other” programs category) were significant covariates on edTPA total scores and edTPA assessment scores when total modified CALI scores, Component 1, and Component 2 scores were predictors. There was one exception in the model containing Component 2 scores predicting edTPA assessment scores where GPA was not significant. The covariates used in these six models accounted for between 44.2% and 53.3% of the variance in scores (i.e., edTPA total scores or assessment scores). However, the most recent published edTPA reports (2016) indicated that demographic covariates only account for 5% of the variance in scores (edTPA Educative Assessment and Meaningful Support, 2016), albeit using different demographic variable coding and excluding the influence of other constructs such as assessment knowledge and confidence.

Exploring no moderation and significant covariates. The lack of a significant moderation effect still produced other key results related to the covariates in the six regression models. Gender was not a significant covariate in any of the six models, but program and GPA were significant in nearly all of the models. Specifically, for programs, AYA and the "Other" programs category were significant in all six models.

A primary point of discussion for this section is related to program. There was a significant relationship between the program that a student was enrolled in and performance on edTPA total and assessment scores. This result largely impacted students in the AYA and "Other" (e.g., Special Education, Music, Art, and Language) programs. The Moderated Multiple Regression results indicated a large negative relationship with the outcome variables for students in AYA and "Other" programs. For example, a student in the AYA program scored -12.80 points less than students in ECED on the edTPA total. For edTPA Assessment scores, the decrease was -5.41 points compared to ECED students. This decrease was also present for students in the Other programs group at -4.02 on edTPA Total scores and -2.36 points for edTPA Assessment scores. Follow-up studies involving interviews and/or focus groups of students and program coordinators may allow for more elaboration and clarification regarding the potential explanations for these differences noted above. Suggestions for future research specific to examining the influence of program in the relationships between assessment knowledge, confidence, and performance are detailed in subsequent sections.

As was presented in the literature review, the knowledge needed by any educator to carry out the process of assessment differs. Models of assessment literacy (Brookhart, 2011; DePascale, 2017; Suskie, 2009) emphasize that the information one teacher might need to answer an assessment-related question is not the same for all teachers. For example, research from Brookhart (2011) highlights that teachers require different assessment-related skills and a deep understand of measurement-related concepts, like percentiles, is not necessarily relevant to all. The findings related to program scores in this study provide some evidence consistent with these theories. Based on Brookhart's (2011) model, the assessment knowledge that an ECED teacher needs would not be equivalent to what a Special Education teacher needs, as one example. Thorough research on the contextual differences between various teaching specializations is necessary in order to interpret these findings but was outside of the scope of this study. Context is considered in the methodological limitations and future directions section of this chapter.

Additionally, GPA was a significant covariate across all six models indicating that a higher GPA resulted in higher scores across assessment knowledge, assessment confidence, and assessment performance. For example, for every one-point increase in GPA, a pre-service teacher's edTPA total score increased by approximately eight points on average. This was also true for edTPA assessment scores with only one GPA point impacting an approximately four-point increase. Interestingly, this indicates that a difference of one point in average cumulative GPA (e.g., having a "B" average instead of a "C" average) is critical for better performance on the edTPA overall and for the assessment domain as well. The implications regarding the importance of cumulative GPA are presented in the following section.

A summary of the major findings presented in the discussion section of this chapter are presented below in Table 50. The focus of this summary is to provide a point of reference for the presentation of implications presented in the following paragraphs. Therefore, Table 50 only summarizes key findings not specifically stated in the research questions and is not a completely exhaustive list of the findings on this study. Minor findings related to demographic and academic information have been excluded. Additionally, findings where no difference or no significant relationship were found are not reported in the table.

Table 50

Summary of Main Findings by Research Question

Research Question (RQ)	Main Findings
RQ 1 and RQ 1A (Pilot Sample)	• Pre-service teachers in the Early Childhood Education (ECED, *n* = 70) and Adolescent Education (AYA, *n* = 45) programs had higher average assessment confidence scores compared to pre-service teachers in Middle Childhood Education (MCED, *n* = 42) and Other programs (*n* = 8). • Average assessment confidence scores differed between pre-service teachers who self-reported they took a course with or without an assessment component. • Average assessment confidence scores differed between pre-service teachers' genders and their mother's highest level of education attained. Pre-service teacher age and average assessment confidence scores were positively related. • Evidence exists for both a unidimensional and a two-component model of pre-service teachers' assessment

	literacy.
RQ 2 (Second Phase Sample)	• Pre-service teachers' Grade Point Averages (GPAs) and average assessment literacy scores were positively related. • Pre-service teachers' average assessment confidence scores and average assessment literacy scores were related. • Pre-service teachers in the Early Childhood Education (ECED, *n* = 34) and Adolescent Education (AYA, *n* = 29) programs had higher average assessment confidence scores compared to pre-service teachers in Middle Childhood Education (MCED, *n* = 14) and Other programs (*n* = 35). • A two-component model of pre-service teachers' assessment literacy was supported, but a unidimensional model is also possible based on cross-loading model modifications and content comparisons across components. • Grade Point Average (GPA) was a positive and significant covariate in almost all of the Moderated Multiple Regression Models predicting total edTPA or edTPA Assessment scores. • Program was a significant covariate in all of the Moderated Multiple Regression models. Group membership in the Adolescent Education (AYA) program followed by the Other program category were predictive of lower edTPA scores and lower edTPA Assessment scores compared to the Early Childhood Education (EDED) program.

Implications

The major implications from the results in this study are discussed for the following groups related to practice and research: (1) Teacher preparation programs and (2) Psychometricians and researchers.

Teacher Preparation Programs

The literature has consistently shown that assessment is an essential component of classroom procedure, and that its proper use can raise students' standards and achievement (Black & Wiliam, 1998; Lukin, Bandalos, Eckhout, & Mickelson, 2004). Classroom assessment, if properly implemented through increased feedback, can improve how well students are learning what is being taught in class, and can also meaningfully boost students' scores on external achievement exams (Black & Wiliam, 1998). However, a popular point of view is that teachers' preparation in assessment and conducting assessment-related activities in the classroom is inadequate (e.g., Fan, Wang, & Wang, 2011; Hills, 1991; Koh, 2011; Plake, 1993). Results from this study describe assessment knowledge in a small sample of pre-service teachers to provide evidence that can contribute to the body of literature on teacher preparation in assessment.

The sample in the current study was at the point of graduation. That is, these students completed the majority or nearly all of their course- and fieldwork, and therefore were at their most educated in their undergraduate program. This group of students, at the culmination of their undergraduate teacher preparation, should (in theory) be most representative of what (and how much) pre-service teachers learn about assessment in their undergraduate programs. Therefore, the temporal placement of the sample used in this study (i.e., in the final months of their teacher preparation program) and the forensic details of their assessment knowledge that were provided can be used to evaluate how teacher preparation programs have impacted student preparedness specifically related to the assessment knowledge on the modified CALI. Reviewing the average level of assessment knowledge from a test external to the program's curriculum with a pre-service teacher sample in their final months of the program can provide criterion related validity to both external measures being considered for use, the standards with which such measures commonly align (i.e., the *Standards*), and the content taught within a program.

Results from this study also have implications for teacher preparation programs related to assessment knowledge and practices measured by assessment literacy instruments, licensure exams, or other high-stakes assessments (e.g., edTPA). Research has shown that both declarative (e.g., paper-and-pencil tests of knowledge) and procedural (e.g., observations, student teaching) knowledge contributes to effective performance in the classroom (Bromme, 2001). However, declarative knowledge is regarded within the psychology literature as the foundation or precursor to procedural knowledge (Anderson, 1982). Thus, a positive relationship is expected between graduating pre-service teachers' knowledge and performance (in general), and as discussed in the literature, specific content knowledge and performance in that content area (e.g., Ball & McDiarmid, 1990; Druva & Anderson, 1983; Minor, Desimone, Lee, & Hochberg, 2016; Robinson, 2017; Whitt & Abigail, 2016). Since items in one component appeared to measure understanding of easier fundamental assessment concepts and items in the second component more difficult applied knowledge, the relationship between knowledge and practice in pre-service teacher assessment need further investigation.

Indeed, there was a moderate, positive correlation between pre-service teachers' knowledge of assessment and their performance based on their edTPA scores (Total EdTPA: $r = .257$, $p = .011$). Interestingly, the magnitude of the correlation was even higher between assessment knowledge and performance specific to assessment (i.e., the

Assessment Domain of the edTPA; $r = .267$, $p = .007$) in the current study. Although basic bivariate correlations, these relationships have implications for teacher preparation programs considering monitoring pre-service teachers' assessment-related knowledge. Observing pre-service teachers' assessment-related knowledge throughout their studies would allow for programs to evaluate any differences between students' edTPA scores and their academic performance should a testing disadvantage arise.

Results related to the influence of GPA on the relationships examined in the current study have implications for not only teacher preparation programs, but for future studies interested in drawing conclusions about predictors of performance-based scores on high-stakes graduation or licensure exams. The significant relationships between cumulative GPA and edTPA Total ($r = .381$, $p < .001$) and edTPA Assessment ($r = .395$, $p < .001$) were positive and moderate to strong. Partial correlations controlling for GPA unsurprisingly rendered the significant relationships between assessment knowledge and edTPA performance nonsignificant ($p > .05$ for all). This finding is supported by a recent meta-analysis that examined the influence of pre-service teachers' test scores, categorized as a basic skills test, professional knowledge exam, content knowledge exam, or the National Teacher's Examination Weighted Uncommon Examinations Total (NTE WCET), and GPA on teaching competence as measured by supervisor/colleague ratings, observations, self-ratings, and student achievement (D'Agostino & Powers, 2009).

Results from the current study indicate that pre-service teacher cumulative GPA has an impact on an external assessment literacy test and performance-based exams. GPA is considered a composite of many factors, and the lack of uniformity in grading practices within and between programs contributes to the conventional assumption that this variable is not predictive of teacher preparedness or performance (D'Agostino & Powers, 2009). In the context of this study related to assessment knowledge and performance, the influence of GPA was apparent. Teacher preparation programs can benefit from this finding by encouraging students' academic achievement and focusing on grades in order to facilitate better outcomes on a high-stakes performance-based exam reflective of preparedness. Thus, teacher preparation programs with high academic standards can rely on their curriculum and internal testing and grading procedures for the data to demonstrate their pre-service teachers' future performance and success in addition to external measures required at the state and/or national level.

Another finding from this study that may have implications for teacher preparation programs addresses assessment knowledge and edTPA performance. The results from this study indicated that scores on the edTPA (Overall) and Assessment edTPA domain in certain programs (e.g., ECED) were significantly higher compared to two other program groups (e.g., AYA and Other). Previous research has indicated that some program-level practices can hinder student performance in teacher preparation programs. In Ledwell and Oyler's (2016) qualitative review of edTPA integration across 19 educators and 12 programs at one university, there were several consequential program-level gatekeeping practices noted such as delaying or denying access to weaker students and counseling students out of teacher education programs. The result of different program performance relative to context and these gatekeeping practices could be used as a point of discussion and reflection within the faculty groups of each program.

The current study showed a relationship between assessment confidence and performance on a traditional measure of assessment literacy, but no significant relationship between assessment confidence and edTPA performance. For teacher preparation programs, this finding means that students' performance on the edTPA may not be impacted by their level of assessment confidence. However, the relationship between confidence and program varied. Table 51 presents the edTPA Total score, edTPA Assessment score, and Average Total Confidence score for each of the four program groups in this study. As stated in previous sections, the edTPA has a total of 75 points possible, and there are 25 points possible on the edTPA Assessment section. Scores for Average Total Confidence had a range of zero to four points (i.e., a the 5-point Likert scale), which was calculated by summing all the Likert ratings divided by 25 items. The sample size for each program relative to the edTPA and confidence scores varies due to missing or incomplete data. The sample sizes are included in the table along with average scores, standard deviations, minimum and maximum scores, and the corresponding medians and interquartile ranges if the data were skewed for each program.

For two of the three main constructs in this study (i.e., assessment confidence, and edTPA performance), a summary of descriptive information by program is provided in Table 51. For example, AYA pre-service teachers had the highest level of confidence in their assessment knowledge, yet compared to students in all other programs, AYA students performed the lowest on the edTPA (Total) and edTPA Assessment. Other descriptive findings from these data include relative similarity between edTPA scores across the ECED, MCED, and Other programs, with only a six-point difference among the three groups. However, pre-service AYA teachers were over eight points below Other programs, which had the next lowest edTPA Total scores. EdTPA Assessment scores showed a similar pattern of performance between the three groups, with ECED and MCED pre-service teachers only separated by less than two points. However, the AYA pre-service teachers scored nearly three points less than the Other program category, which had the second lowest scores. These results highlight the differences between three of the program groups and the AYA program group regarding edTPA performance and assessment confidence.

Table 51

Summary of edTPA Total, edTPA Assessment, and Assessment Confidence Scores by Program

Variable	Program	*n*	*M* (*Mdn*)	*SD* (IQR)	Min/Max
edTPA Total (*N* = 97)	ECED	33	46.76	5.66	38/62
	MCED	13	43.38	9.26	26/61
	AYA	21	32.43	6.79	19/50
	Other	30	40.83	6.48	29/54
edTPA Assessment (*N* = 101)	ECED	34	15.78	3.07	5/21
	MCED	13	14.31	4.29	6/21
	AYA	21	9.57	3.70	4/19
	Other	33	12.38	3.33	6/18
Assessment Confidence (*N* = 112)	ECED	34	2.71 (2.88)	.59 (.62)	.80/3.52
	MCED	14	2.66	.59	1.36/3.32
	AYA	29	2.83 (2.96)	.54 (.52)	.56/3.40
	Other	35	2.49	.48	1.72/3.52

Note. ECED = Early Childhood Education Program, MCED = Middle Childhood Education Program, AYA = Adolescent Education Program, Other = Other Education Programs. *Mdn* = Median. IQR = Interquartile Range.

Other measures of assessment knowledge should also be considered. One example of an alternative assessment that would allow pre-service teachers to demonstrate assessment knowledge is project-based learning. Frank and Barzilai (2004) conducted a study with pre-service science teachers who created a portfolio, written reports, presentations, and physical models of classroom products. Such projects related to assessment could include giving mock score cards for a common state assessment to each pre-service teacher and their goal would be to prepare for and execute a short conversation/presentation that clarifies the results. Brookhart (1999) discusses several alternative practices for pre-service teachers who need to learn the communication of assessment results and grading. For instance, group discussions on topics like the relationship between the type of instruction and the approach to grading, and critiquing rubrics are suggested. Additionally, Brookhart (1999) suggests providing pre-service teachers with different scenarios where they have to produce a series of good and bad solutions for topics such as assigning zeros or working with percentages.

Implications for teacher education programs also extend to the national level within the context of legislation like the Every Student Succeeds Act (ESSA). Because of the influence of such acts, educators are faced with the centrality of assessment and accountability in measuring not only student academic improvement, but also teacher and institutional effectiveness (DeLuca & Bellara, 2013). With the current and previous legislation (e.g., the No Child Left Behind [NCLB] act), the frequency and duration of standardized assessment has increased across the country. The ESSA also restructured accountability systems at the local and state level. For example, the weight of district and state tests impacts the score or report card assigned to a school. Accountability systems also often include measures like student growth percentiles, which are used to reflect teacher performance. In turn, using student assessment data as measures of teacher effectiveness, and consequently school and district effectiveness, emphasizes the need to create assessment literate teachers who are equipped to interact with assessment in order to understand the implications of its use.

In the context of the state of Ohio and its policies, the need to create assessment literate teachers is apparent. Ohio's 2017 newly-approved accountability system is comprised of five indicators. Three of these five indicators are related to specific outcomes measured by state-required tests (Ohio Department of Education [ODE]). For example, according to the ODE's website, the achievement indicator, selected by the state as a measure of performance and growth, is evaluated by using statewide English Language Arts and Math assessment scores from primary and secondary schools. The results from these assessments contribute to Value-Added Modeling (VAM) or student growth model scores, which attempt to isolate the contribution of a teacher or school from their students' performance (McCaffrey, Lockwood, Koretz, & Hamilton, 2003).

VAMs have many forms with a similar core premise – using prior student performance over time (e.g., two test administrations are one year or across years) in an attempt to measure a teacher's impact on student achievement (for an overview of models used see Bassiri, 2015). Value-added score reports typically present teachers with several assessment-related concepts such as averages, class size, average percentile, average predicted scores, average predicted percentiles, a growth measure score, standard errors, and school versus reference group averages (ODE EVAAS Report, 2015). The range of assessment-related terminology and concepts reported from VAMs is one example of why teachers need more training to become assessment literature in the classroom and perhaps beyond. The complexity of VAM, which uses student assessment data to gauge teacher effectiveness, can be daunting. Teachers may not need to understand every statistical concept behind the VAM used by their district or at the state and national levels, but teachers should be equipped with enough assessment knowledge to interpret their VAM score, as well as its implications.

To better prepare pre-service teachers for educational policy and procedures related to assessment, course time focusing on the current climate of the education system could be bolstered. This course time could cover topics such as VAM and the political context outlined above. In the current study, pre-service teachers responded with approximately 48% accuracy on Component 2 items. The broad characteristics of the content contained in Component 2 items, outlined in Chapter 4, include assessment concepts such as scoring/grading, communicating assessment results, and standardized testing ethics/procedures. These broad themes relate more the topics surrounding national conversations on student testing and teacher evaluation or performance scores.

A sample of curricula from the teacher preparation program in this study revealed one three-credit course is required either in the first, second, or third semester of course work for ECED, MCED, and AYA pre-service teachers. In reviewing publicly-available syllabi, this course had four objectives, and one of these objectives was to understand the economic, legal, and political context of schools. In addition to this course objective (i.e., the political context of schools), regular seminars or workshops could be provided to keep students informed of policy changes and introduce them to concepts related to assessment at the district, state, and federal levels. Therefore, exposing pre-service teachers to these more difficult applied concepts (i.e., Component 2) across the curriculum may help to demonstrate assessment literacy's relevance beyond the classroom to the greater educational policy climate, and potentially encourage a less superficial understanding of assessment via application across courses.

Psychometricians and Researchers

There is a continued need for assessment-related teacher preparation research to incorporate advanced statistical techniques in analyzing data. This is of particular importance to studies that investigate populations of pre-service teachers, their preparation, and the contributions of the teacher preparation program and curricula. Existing research has predominately used basic descriptive statistics, univariate inferential tests, mixed-methods approached, and qualitative analyses, with a smaller proportion using multivariate inferential tests (DeLuca & Klinger, 2010; Mertler & Campbell, 2005; Smith, Worsfold, Davies, Fisher, & McPhail, 2013). The current study used Rasch Analysis, Rasch Principal Components Analysis (PCA), Confirmatory Factor Analysis (CFA), and Moderated Multiple Regression models, which have rarely if ever been used to analyze data in this area of research. For instance, by using Rasch analysis to investigate the items, persons, and structure of the CALI, this study provided information on specific areas of assessment knowledge that cannot be appraised with other statistical techniques in measurement (e.g., item difficulty and person ability alignment).

One psychometric consideration for researchers is the use of Rasch Analysis to detect dimensionality and identifying underlying internal structures. Rasch PCA analyses compute general “Components” comprised of items that are related, and do not present set “Factors” of items that represent latent variables (Linacre, 1998). Brentari and Golia (2007) and Brentari, Golia, and Manisera (2007) describe the concept of dimensionality as a continuum and therefore even though statistical information may suggest multidimensionality, the dimensions exist on a continuum. The Rasch models used in this study provided evidence for both a unidimensional and multi-dimensional measure. Several considerations are addressed by Linacre (2012) who stated that multidimensionality exists to some extent in all measures. It is up to the researcher to decide if the multi-dimensionality merits dividing the items into separate tests. This consideration should be made carefully in borderline cases such as in the current study where evidence for a complete multi-component measure was inconclusive. It may be that a select few items are off-dimension or unusual and should be eliminated from the measure. For example, the modified CALI included two ethics questions which both loaded on the same component.

It is important to subject measures such as the CALI to more intense psychometric scrutiny. Prior to this study, researchers using the CALI overlooked the scant quantitative validity evidence, and overwhelmingly did not question the internal structure. Since the development of the CALI and other assessment literacy measures, researchers have blindly used the *Standards*-based internal structure without any factor analytic support. Although the current study cannot draw any definitive conclusions in support of one internal structure, results from this study at least demonstrate that the "clean" and "tidy" *Standards*-based conceptualization of assessment knowledge is questionable, and perfect alignment with the seven *Standards* is highly improbable regardless of the sample used.

Furthermore, the onus is also on the academic research community to consider and promote the importance of measurement as a formative necessity prior to (and while) conducting any study. More often than not, measurement is an afterthought in "academic" research, whereas evidencing reliability and validity should always be the first step in planning, developing, or piloting any study. In addition, providing quality psychometric evidence (i.e., reliability and validity) through measurement should always be the first consideration when piloting or using a measure in a new context. Previous use (i.e., either unguided or misinformed) of assessment literacy measures in research is evidence of this disregard for or misunderstanding of measurement among academics and other researchers. For example, the CALI was originally used in a sample of both pre-service and in-service teacher, and the psychometric results demonstrated lower internal consistency reliability (KR-20 = .54) for in-service teachers (Mertler, 2003). The measure continued to be developed, however, only in relation to pre-service teachers and the in-service teacher sample reliability, or the decision to exclude this population, was not addressed (Mertler, 2005). The development of this measure has not updated in over ten years, and yet, the CALI is still being used to measure assessment literacy in both populations (i.e., pre-service and in-service teachers) when psychometric evidence is lacking or patently incorrect. This is just one such example of an assessment literacy measure that requires additional development.

The last implication discussed here is the use of the modified CALI as a measure of assessment literacy. Two main points must be noted in relation to the suggested use of the modified CALI. First, both the pilot and phase two samples performed only slightly above chance (i.e., 50%). This evidence suggests that the overall measure was too difficult for the sample, with nearly seven questions not answered correctly by anyone during the first administration (i.e., one person answered three of these correctly). Therefore, continued adaption of the modified CALI or a similar measure of assessment literacy specific to pre-service teachers is suggested. If the edTPA were to be issued as an official licensure exam, teacher preparation educators could create a meaningful measure of assessment literacy that prepares students and programs for the edTPA. The edTPA was created to align with state and national standards such as the Common Core State Standards and the Interstate Teacher Assessment and Support Consortium (InTASC). These standards should be a starting point for the creation of an edTPA-specific assessment literacy measure.

Limitations

Three sections of limitation are presented below: (1) Conceptual, (2) Methodological, and (3) Statistical and Psychometric. These limitations are accompanied by suggestions for future research.

Conceptual Limitations and Future Directions

One conceptual limitation to the current study is the use of the newly-conceptualized and developed measure of assessment confidence. Assessment has been integrated into teacher confidence research as a component of pedagogical knowledge, basic skills (e.g., classroom management), or specific arears of content knowledge (e.g., math, science, language arts). However, sections of this study's literature review highlighted the need for research in this area such as assessment confidence. Assessment confidence may be related to specific areas of teaching (i.e., science teacher's assessment confidence), but it may also be a part of the foundation of basic teaching skills that require assessment. Therefore, research on this construct is needed in order to determine the role that it might play within assessment and other confidences related to teaching.

To further explore the concept of assessment confidence, practitioners and researchers can incorporate measures of confidence into research studies and program evaluation. The current study did not find confidence as a moderator of the assessment knowledge and performance relationship. However, there was a significant and strong relationship between assessment confidence and pre-service teacher scores on a more traditional measure of assessment knowledge (i.e., the CALI). The finding that assessment confidence was strongly and positively related to assessment knowledge may indicate the need for more confidence research in that specific context (i.e., traditional formal or informal multiple-choice knowledge tests). Many states still use traditional assessments like the PRAXIS II or state specific tests like the Ohio Assessment for Educators as a road to licensure (Educational Testing Services PRAXIS State Requirements). By monitoring student confidence through student self-reported measures or one-on-one interviews, programs can mitigate the effects of low confidence on such tests as the PRAXIS as well as other course-level exams.

Methodological Limitations and Future Directions

A methodological disadvantage of this study is the use of a sample with high dependency on its context. These students, the education, assessment preparation, and field experience they received are dependent upon the curricula of the university and even the requirements of the state. Any sample of teacher education students is highly specific to its context as all teacher preparation programs are different. This study in particular investigates several teacher education preparation programs within one university, ranging from Early Childhood Education (ECED), Middle Childhood Education (MCED), Adolescent Childhood Education (AYA), and Special Education to other specialties in other colleges within the university such as Music, Art, Physical Education, and Foreign Language teaching. Students were therefore studying under different programs with varied curricula and pedagogies even within this sample. These curricular differences could impact generalizing results to other preparation programs, which may or may not focus on assessment in different contexts (i.e., different kinds of assessment such as Art and Music) and in greater or lesser capacities.

This specific influence of curricular differences was noticed in the students' self-reported attendance in courses with assessment, courses that focused only on assessment, attending a workshop that had an assessment component, and a workshop that emphasized only assessment. For example, reviewing the average level of assessment confidence for students who reported taking versus not taking an assessment-specific course indicated that those who did not take an assessment-specific course were more confident in their assessment knowledge. The same was true for assessment knowledge where students who did not take a course that was assessment-specific had a higher average score on the modified CALI than students who did take an assessment-specific course. However, the group sizes for the attending and not attending groups were drastically different. Consequently, this spurious finding likely stemmed from very unequal sample sizes and therefore is also a statistical limitation.

The above methodological concern was clear in the current study as the results indicated significant differences in performance according to program. Future research can use the CALI to investigate pre-service teacher assessment knowledge and assessment confidence across a more homogenous sample of teacher education students at several universities. For instance, obtaining a larger sample of just ECED students across multiple universities throughout the state of Ohio would yield different results. It would also provide a more program-specific set of outcomes related to assessment knowledge and confidence.

Alternatively, future studies could incorporate a mixed-methods approach to the research questions asked in this study. One potential goal from a future mixed-methods study could be to triangulate the quantitative findings with a series of interviews and/or small group discussions and observations. Triangulation is a process used by qualitative researchers to provide evidence of validity by analyzing multiple data sources for inconsistencies (Merriam & Tisdell, 2015). This approach could use data in the form of edTPA scores, a measure of assessment knowledge such as the modified CALI, a measure of assessment confidence or confidence in general, observations, and interviews in order to investigate assessment knowledge and preparation with more depth. The interviews would target a series of stakeholders engaged in pre-service teacher preparation. Interviewees may include, but are not limited to, students (pre-service teachers), professors, administrators, in-service teachers who participate in student teaching programs, researchers, and local- or state-level education officials. Interview questions could topically be organized by the main constructs in this study (i.e., assessment knowledge, assessment confidence, and performance on a portfolio-based measure of teacher preparedness). The goal of these interviews would be to gather deeper knowledge and understanding on pre-service teacher assessment preparation, confidence, and edTPA use across all invested parties.

Additionally, observations of lectures attended by the participants could be helpful in explaining differences between programs. One way these observations could be conducted is by observing the same course (i.e., the same curriculum) across different sections (i.e., programs). Ideally this undertaking would span several semesters in order to cover several course sections and professors. Observations of the graduating students as they are student teaching could also be implemented. This observation would specifically monitor the engagement of the pre-service teacher with assessment related activities. This information could then be combined with the quantitative data strategies used in the present study, along with other written documents like curriculum and syllabi.

Another methodological limitation from the current study is using one measure of assessment literacy for all programs and students, regardless of their experience level or pedagogical specialization (e.g., ECED versus AYA, second- and third-year versus graduating pre-service teachers). The results from this study indicated that each group of pre-service teachers has differing assessment knowledge (and confidence and performance) at the program level. Previous research has used the same version of the CALI for both pre-service and in-service teachers and also reported differences between groups (Mertler, 2003). As the body of literature related to teacher assessment literacy continues to demonstrate this increasingly common finding, the universal use of the CALI considerably weakens statistical conclusion validity, and ultimately the substantive conclusions drawn, in this area of research. Therefore, one measure of assessment literacy may be methodologically (and psychometrically) inappropriate.

Future researchers should consider the methodological (and measurement-related) consequences of using the CALI with populations and in contexts for which little or inaccurate validity evidence exists. The validation process should answer questions of "…validity for what or in what context/for what purposes?" or "…validity for whom or in which population?" This study provides evidence of variability in assessment knowledge in different contexts (and for different purposes) and between populations. For example, researchers must appraise the target population's characteristics and the context before administering any measure of assessment literacy. As an example from the current study, the differences in assessment knowledge between ECED and AYA pre-service teachers should be a fundamental validity concern prior to conducting more research in this area. Future research should consider creating multiple versions of the

CALI for teacher education program specializations and different versions for pre-service and in-service teachers.

Statistical and Psychometric Limitations and Future Directions

The first phase of this study included responses from 165 participants. These responses were used to investigate the first objective of this study, which was to analyze the psychometric properties (i.e., reliability and validity) of the modified CALI. This sample size was not large enough to conduct Exploratory Factor Analysis (EFA). EFA allows for the investigation of theoretical constructs, or factors, which might be represented by a set of items (Tabachnick & Fidell, 2001). EFA is used when researchers have no predetermined hypotheses or prior theory about the nature of the underlying factor structure of their measure. It is an inductive approach using factor loadings to uncover the structure of the data. Since EFA is exploratory by nature, no inferential statistical processes are used. Costello and Osborne (2005) caution researchers that EFA is a "large-sample" procedure. In fact, Tabachnick and Fidell (2001) suggest a study have at least 300 participants for this type of analysis. Generalizable or replicable results are unlikely if the sample is too small. In the case of this study, 165 participants did not meet the suggested 300 cases proposed by Tabachnick and Fidell (2001). Therefore, EFA was not used to analyze the data in Phase 1 of this study. Future research should obtain a sample large enough for EFA if this measure is used in a new population. As was shown by the current study, despite alignment between a measure's content and a set of established standards, the underlying structure of the measure may not be consistent with that of external standards.

The use of CFA as a way to confirm the results of the Rasch PCA is also a statistical limitation. A major difference between CFA and Rasch PCA is the transformation of data. Rasch PCA depends on linear transformations of the data using residuals which are forced to be orthogonal (Grimby, Tennant, & Tesio, 2012). Rasch PCA also does not proportion variance like CFA in that CFA takes into account the correlations between all possible variables (Kim, 2008). Since the structure of the data in these two models differs, it is difficult to make comparisons between the results (i.e., if one can confirm the results in the other). One method that can be used to achieve some level of comparably in the structure of the data includes using all the Rasch estimates with Diagonally Weighted Least Squares (DWLS) estimation in CFA or in Principle Axis Factoring (Li, 2016). Principle Axis Factoring partitions variance into shared and unique, which is consistent with the process of CFA. It would also prove to be a more comparable and appropriate companion to a CFA analysis than Rasch PCA.

Additionally, decisions made in the confirmatory phase of this study (i.e., the CFA procedures) present statistical limitations such as the use of the ULS estimation method as well as parceling. Due to the small sample size in the confirmatory study (N = 112), the ULS estimation method was one of a few limited options (Jung, 2013). The statistical limitation associated with ULS of most concern is its reduced error bias (Obenchain, 1975). Secondly, the use of parceling is controversial in the field of measurement. When items are created into parcels, item level information can be distorted (Bandalos, 2002). In this study, parcels were used to investigate content level information and not item level details. However, by parceling items the model may not have the same accuracy in representing the data. Future research should obtain a larger sample and compare the parceling approach to the item-level approach.

General limitations related to survey research, such as global ratings and self-report scales, must also be considered. Global ratings refer to the overall impression or summary statement of the construct of interest (Kazdin, 2003). In this study, the measure investigated assessment confidence. Although global ratings provide a very flexible and convenient assessment format for soliciting judgments, they are not without consequences. Global ratings tend to be general and lack the sensitivity needed to access the construct of interest. In the case of assessment confidence, clearly defining what it means to be “Completely,” “Mostly,” or “Neither Confident or Unconfident” would be beneficial. However, this does not remove the initial evaluation on the part of the respondent when he/she reads the word and creates his/her own idea of what is means to be “Completely” confident, for example. Future studies should examine the use of different response scales for comparison purposes and determine which response scale or structure is best for the population of interest. Choosing a report scale specific to the population of study and briefing the population on the meaning of the scale may lead to a more accurate report of constructs like confidence.

Participants used self-report to answer items on the modified CALI, asking if they had taken any assessment-related coursework, had classroom experience, and about their overall preparation. These items asked students if they had or had not taken a course with an assessment component or a course which had an assessment-only focus. Based on the results, in some cases students may have misinterpreted the question. For example, when the second phase sample was asked if they had student teaching experience, 4.5% (*n* = 5) of the sample responded “No” when the graduation requirements clearly include student teaching. Participants also were not asked to list the specific names of courses taken with assessment components or with assessment-only foci. Therefore, the data from this study do not contain enough detail to corroborate evidence of taking assessment-related courses and if they were actually attended by the participants.

Self-report measures or scales that require individuals to report on aspects of their own personality, emotions, cognitions, or behaviors, are also problematic (Kazdin, 2003). Although there are practical benefits to using self-report measures, this mode of assessment is subject to social desirability, with the confidence scale and academic components on the CALI being no different. The possibility of bias and distortion on the part of the subjects in light of their own motives, self-interest, or to “look good” is elevated with this type of measure. This is also related to the obtrusiveness of the measure. The construct of interest measured in the CALI and the added confidence component are obtrusive and have high face validity. Although the directions indicated that these results would not be shared with anyone specifically, and that they would be reported collectively, participants may still have been worried about the perception of others and responded more favorably. Future studies should consider deviating from this method of assessment and perhaps use direct observation or other measures.

Lastly, the methodology of the Rasch model and its role in the psychometric and factor analytic evaluation of this measure must also be discussed. The Rasch model is grounded in its use of assumptions, which are more difficult to meet than those of

Classical Test Theory (CTT) (Wright & Stone, 1979). However, according to the Rasch model, when data do not adequately fit the model, the instrument construction process must be reiterated. The need for a total re-creation of a measure could occur if the items in questions are poorly constructed, incomprehensible to the sample, or if the respondent group's abilities and the difficulty of the items are misaligned. This limitation must be mentioned due to the relationship between items and persons that the Rasch model evaluates. The refinement of the modified CALI was conducted based on the responses unique to this sample. Future research should focus on defining the construct, especially assessment confidence, validating the scores on the revised measure, and eventually demonstrate the potential flexibility of the measure with other teacher education programs.

Conclusion

Assessment literacy is gaining attention across the U.S. as the use of assessment data and results continue to evolve both in and out of the classroom. Teachers, and therefore teacher education programs, are at the forefront of this change as they are forced to rapidly adapt to policy changes and licensure expectations. Presently, measures of assessment literacy have been underutilized in pre-teacher teacher populations. The current study investigated the psychometric properties (i.e., reliability and validity) of the modified Classroom Assessment Literacy Inventory (CALI) – a measure of teacher assessment literacy – using a sample of pre-service teachers in their final semester of study. A confidence scale was included to examine psychometric construct evidence of assessment confidence. Additionally, this psychometric analysis examined the component structure of the modified CALI in order to determine if any distinct domains of knowledge were present in this sample. Lastly, the relationships between pre-service teacher assessment knowledge, assessment confidence, and performance on a portfolio-based measure of teacher preparedness were explored.

Results indicated the possible presence of two underlying dimensions of assessment literacy in this sample of pre-service teachers, and the likelihood of one unidimensional component representing assessment literacy in pre-service teachers. The evidence supporting a multi-dimension construct was inconclusive and in order to definitely determine the existence of a second component, replication of this study is necessary. Confidence positively influenced participants' scores on each of these components and the total CALI scores. Additionally, the two possible components of knowledge on the modified CALI had a significant positive relationship with edTPA assessment scores. While confidence only influenced performance on a traditional paper-and-pen measure of assessment literacy (i.e., the modified CALI), the knowledge assessed by these two components did have a relationship with portfolio-based performance (i.e., the edTPA). Moreover, differences between pre-service teacher education programs' performance at the same university were noted. These results provided meaningful information about measuring assessment knowledge and confidence and preparing pre-service teachers to take exams with assessment related sections. This study added to the growing literature base surrounding teacher assessment literacy. Teacher education programs impacted by the implementation of performance-based assessments like the edTPA, should interpret these exploratory results with caution, but can consider the use of a measure of assessment literacy as a means of monitoring the progression of pre-service teacher assessment knowledge.

APPENDICES

APPENDIX A

MODIFIED CLASSROOM ASSESSMENT LITERACY INVENTORY (CALI)

Appendix A
Modified Classroom Assessment Literacy Inventory (CALI)
(Administered via Qualtrics with 35 Multiple-Choice Items and 35 Confidence Items)

Hello, my name is Kelli Ryan and I am a Ph.D. Candidate interested in researching educator familiarity about assessment. Educators vary widely in their knowledge of assessment, and I am trying to understand what concepts are understood and how confident educators are in their assessment knowledge. If you agree to participate, you will be asked to complete a multiple-choice survey. Please set aside at least 30 minutes to complete this survey. These data will help us understand what teachers know about assessment and how confident they are in your assessment knowledge.

All responses to this survey are anonymous and will not be linked to your name or any other identifying information. The data will be stored along with randomly assigned ID numbers that are also not linked to or stored with your names or any other identifying information. Please find a copy of the entire consent document attached here (Live link to PDF of IRB # 16-747 informed consent). You may stop participating at any time. If you do not wish to participate, please exit this screen now. By clicking continue you acknowledge you consent to participate.

2.2 What is your age in years to the nearest whole number?

2.3 With what gender do you identify?
○ Male
○ Female
○ Other (please specify) ____________________

2.4 What is your race?
○ Caucasian or Non-Hispanic
○ Black or African American
○ Hispanic or Latino
○ American Indian or Alaska Native
○ Asian
○ Native Hawaiian or Pacific Islander
○ Bi-Racial or Multi-Racial
○ Other (please specify) ____________________

2.8 What is your current cumulative undergraduate grade point average (i.e., GPA) on a 4.0 scale?

2.9 What is your current student status according to university credit hour requirements?
○ Freshman
○ Sophomore

❍ Junior
❍ Senior
❍ Other (please specify) ____________________
2.10 Are you a first-generation college student?
❍ Yes
❍ No

2.11 What is your mother's highest level of education completed?
❍ Some High School, No Diploma
❍ High School Diploma or General Education Development (GED)
❍ Technical Diploma or Degree
❍ Some College, No Degree
❍ Associate's Degree
❍ Bachelor's Degree
❍ Master's Degree
❍ Doctoral Degree (PhD) or Medical Degree (MD)
❍ Professional Degree (e.g., Lawyer, Dentist, Optometrist)
❍ Other (please specify) ____________________

2.12 What is your father's highest level of education completed?
❍ Some High School, No Diploma
❍ High School Diploma or General Education Development (GED)
❍ Technical Diploma or Degree
❍ Some College, No Degree
❍ Associate's Degree
❍ Bachelor's Degree
❍ Master's Degree
❍ Doctoral Degree (PhD) or Medical Degree (MD)
❍ Professional Degree (e.g., Lawyer, Dentist, Optometrist)
❍ Other (please specify) ____________________

The following questions ask about your experience with assessment. Some examples of assessment experience are writing test items, analyzing student performance, administering standardized assessments, and communicating scores.

2.13 Have you ever taken a course (a few weeks or more) in which the topic was only assessment?
❍ Yes
❍ No

2.14 If you have taken a course (a few weeks or more) in which the topic was only assessment, how course many you taken? Please use whole numbers.

2.15 Have you ever taken a workshop (a few days or less) in which the topic was only

assessment?
○ Yes
○ No

2.16 If you have taken a workshop (a few days or less) in which the topic was only assessment, how many workshops many you taken? Please use whole numbers.

2.17 Have you ever taken a course (a few weeks or more) in which assessment was one of multiple topics covered?
○ Yes
○ No

2.18 If you have taken a course (a few weeks or more) in which assessment was one of multiple topics, how many courses have many you taken? Please use whole numbers.

2.19 Have you ever taken a workshop (a few days or less) in which assessment was one of multiple topics covered?
○ Yes
○ No

2.20 If you have taken a workshop (a few days or less) in which assessment was one of multiple topics, how many workshops have many you taken? Please use whole numbers.

2.21 In your undergraduate program, did/do you have experience in the classroom in any of the following capacities? Please select all that apply.
□ Observer
□ Teacher's Assistant
□ Student Teacher (e.g., teaching with supervision)
□ Lead Teacher (e.g., teaching without supervision)
□ Current Licensed Teacher
□ Other please specify (i.e., part time job) ____________________

2.22 Which of the following best describes your perception of the level of preparation for the overall job of being a classroom teacher that resulted from your undergraduate teacher preparation program?
○ Very Unprepared
○ Somewhat Unprepared
○ Somewhat Prepared
○ Very Prepared

2.23 Which of the following best describes your perception of the level of preparation for assessing student performance that resulted from your undergraduate teacher preparation program?

- ❍ Very Unprepared
- ❍ Somewhat Unprepared
- ❍ Somewhat Prepared
- ❍ Very Prepared

3.1 You will now be asked 35 multiple choice questions about assessment. Choose the best answer to each question. Even if you are not sure of your choice, mark that response. After you answer each question to the best of your ability, you will rate how confident you are in your answer. [Adapted from the Classroom Literacy Assessment Inventory (n.d.), by C. Mertler, Bowling Green State University]

3.2 What is the most important consideration in choosing a method for assessing student achievement?

- ❍ The ease of scoring the assessment
- ❍ The ease of preparing the assessment
- ❍ The accuracy of assessing whether or not instructional objectives were attained
- ❍ The acceptance by the school administration

3.3 How confident are you in your answer to the above question?

- ❍ Completely Unconfident
- ❍ Mostly Unconfident
- ❍ Neither Confident nor Unconfident
- ❍ Mostly Confident
- ❍ Completely Confident

3.4 When scores from a standardized test are said to be “reliable,” what does it imply?

- ❍ Student scores from the test can be used for a large number of educational decisions.
- ❍ If a student retook the same test, he or she would get a similar score on each retake.
- ❍ The test score is a more valid measure than teacher judgments.
- ❍ The test score accurately reflects the content of what was taught.

3.5 How confident are you in your answer to the above question?

- ❍ Completely Unconfident
- ❍ Mostly Unconfident
- ❍ Neither Confident nor Unconfident
- ❍ Mostly Confident
- ❍ Completely Confident

3.6 Mrs. Bruce wished to assess her students' understanding of the method of problem solving she had been teaching. Which assessment strategy below would be most valid?

- ❍ Select a textbook that has a "teacher's guide" with a test developed by the authors.
- ❍ Develop an assessment consistent with an outline of what she has actually taught in class.
- ❍ Select a standardized test that provides a score on problem solving skills.

- ○ Select an instrument that measures students' attitudes about problem solving strategies.

3.7 How confident are you in your answer to the above question?

- ○ Completely Unconfident
- ○ Mostly Unconfident
- ○ Neither Confident nor Unconfident
- ○ Mostly Confident
- ○ Completely Confident

3.8 What is the most effective use a teacher can make of an assessment that requires students to show their work (e.g., the way they arrived at a solution to a problem or the logic used to arrive at a conclusion)?

- ○ Assigning grades for a unit of instruction on problem solving.
- ○ Providing instructional feedback to individual students.
- ○ Motivating students to attempt innovative ways to solve problems.
- ○ None of the above.

3.9 How confident are you in your answer to the above question?

- ○ Completely Unconfident
- ○ Mostly Unconfident
- ○ Neither Confident nor Unconfident
- ○ Mostly Confident
- ○ Completely Confident

3.10 Ms. Green, the principal, was evaluating the teaching performance of Mr. Williams, the fourth-grade teacher. One of the things Ms. Green wanted to learn was if the students were being encouraged to use higher order thinking skills in the class. What documentation would be the most valid to help Ms. Green to make this decision?

- ○ Mr. Williams’ lesson plans.
- ○ The state curriculum guides for fourth grade.
- ○ Copies of Mr. Williams’ unit tests or assessment strategies used to assign grades.
- ○ Worksheets completed by Mr. Williams’ students, but not used for grading.

3.11 How confident are you in your answer to the above question?

- ○ Completely Unconfident
- ○ Mostly Unconfident
- ○ Neither Confident nor Unconfident
- ○ Mostly Confident
- ○ Completely Confident

3.12 A teacher wants to document the validity of the scores from a classroom assessment strategy she plans to use for assigning grades on a class unit. What kind of information would provide the best evidence for this purpose?

- ○ Have other teachers judge whether the assessment strategy covers what was taught.

❍ Match an outline of the instructional content to the content of the actual assessment.
❍ Let students in the class indicate if they thought the assessment was valid.
❍ Ask parents if the assessment reflects important learning outcomes.

3.13 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.14 Which of the following would most likely increase the reliability of Mrs. Lockwood's multiple-choice end-of-unit examination in physical science?
❍ Use a blueprint to develop the test questions.
❍ Change the test format to true-false questions.
❍ Add more items like those already on the test.
❍ Add an essay component.

3.15 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.16 Ms. Gregory wants to assess her students' skills in organizing ideas rather than just repeating facts. Which words should she use in formulating essay exercises to achieve this goal?
❍ compare, contrast, criticize
❍ identify, specify, list
❍ order, match, select
❍ define, recall, restate

3.17 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.18 Mr. Woodruff wanted his students to appreciate the literary works of Edgar Allen Poe. Which of his test items shown below will best measure his instructional goal?
❍ "Spoke the raven, nevermore." comes from which of Poe's works?
❍ True or False: Poe was an orphan and never knew his biological parents.

- ○ Edgar Allen Poe wrote: 1. Novels 2. Short Stories 3. Poems 4. All of the above.
- ○ Discuss briefly your view of Poe's contribution to American literature.

3.19 How confident are you in your answer to the above question?

- ○ Completely Unconfident
- ○ Mostly Unconfident
- ○ Neither Confident nor Unconfident
- ○ Mostly Confident
- ○ Completely Confident

3.20 Several students in Ms. Atwell's class received low scores on her end-of-unit test covering multi-step story problems in mathematics. She wanted to know which students were having similar problems so she could group them for instruction. Which assessment strategy would be best for her to use for grouping students?

- ○ Use the test provided in the "teacher's guide."
- ○ Have the students take a test that has separate items for each step of the process.
- ○ Look at the student's records and standardized test scores to see which topics the students had not performed well on previously.
- ○ Give students story problems to complete and have them show their work.

3.21 How confident are you in your answer to the above question?

- ○ Completely Unconfident
- ○ Mostly Unconfident
- ○ Neither Confident nor Unconfident
- ○ Mostly Confident
- ○ Completely Confident

3.22 Many teachers score classroom tests using a 100-point percent correct scale. In general, what does a student's score of 90 on such a scale mean?

- ○ The student answered 90% of the items on this test correctly.
- ○ The student knows 90% of the instructional content of the unit covered by this test.
- ○ The student scored higher than 90% of all the students who took the test.
- ○ The student scored 90% higher than the average student in the class.

3.23 How confident are you in your answer to the above question?

- ○ Completely Unconfident
- ○ Mostly Unconfident
- ○ Neither Confident nor Unconfident
- ○ Mostly Confident
- ○ Completely Confident

3.24 Students in Mr. Jakman's science class are required to develop a model of the solar system as part of their end-of-unit grade. Which scoring procedure below will maximize the objectivity of assessing these student projects?

❍ When the models are turned in, Mr. Jakman identifies the most attractive models and gives them the highest grades, the next most attractive get a lower grade and so on.
❍ Mr. Jakman asks other teachers in the building to rate each project on a 5-point scale based on their quality.
❍ Before the projects are turned in, Mr. Jakman constructs a scoring key based on the critical features of the projects as identified by the highest performing students in the class.
❍ Before the projects are turned in, Mr. Jakman prepares a model or blueprint of the critical features of the product and assigns scoring weights to these features. The models with the highest scores receive the highest grade.

3.25 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.26 At the close of the first month of school, Mrs. Friend gives her fifth grade students a test she developed in social studies. Her test is modeled after a standardized social studies test. It presents passages and then asks questions related to understanding and problem definition. When the test was scored, she noticed that two of her students—who had been performing well in their class assignments—scored much lower than other students. Which of the following types of additional information which would be most helpful in interpreting the results of this test?
❍ The gender of the students.
❍ The age of the students.
❍ Reliability data for the standardized social studies test she used as the model.
❍ Reading comprehension scores for the students.

3.27 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.28 Frank, a beginning fifth grader, received a G. E. (grade equivalent score) of 8.0 on the Reading Comprehension subtest of a standardized test. This score should be interpreted to mean that Frank
❍ can read and understand 8th grade reading level material.
❍ scored as well as a typical beginning 8th grader scored on this test.

○ is performing in Reading Comprehension at the 8th grade level.
○ will probably reach maximum performance in Reading Comprehension at the beginning of the 8th grade.

3.29 How confident are you in your answer to the above question?
○ Completely Unconfident
○ Mostly Unconfident
○ Neither Confident nor Unconfident
○ Mostly Confident
○ Completely Confident

3.30 When the directions indicate each section of a standardized test is timed separately, which of the following is acceptable test-taking behavior?
○ John finishes the vocabulary section early; he then rechecks many of his answers in that section.
○ Mary finishes the vocabulary section early; she checks her answers on the previous test section.
○ Jane finishes the vocabulary section early; she looks ahead at the next test section but does not mark her answer sheet for any of those items.
○ Bob did not finish the vocabulary section; he continues to work on that section when the testing time is up.

3.31 How confident are you in your answer to the above question?
○ Completely Unconfident
○ Mostly Unconfident
○ Neither Confident nor Unconfident
○ Mostly Confident
○ Completely Confident

3.32 Ms. Camp is starting a new semester with a factoring unit in her Algebra I class. Before beginning the unit, she gives her students a test on the commutative, associative, and distributive properties of addition and multiplication. Which of the following is the most likely reason she gives this test to her students?
○ The principal needs to report the results of this assessment to the state testing director.
○ Ms. Camp wants to give the students practice in taking tests early in the semester.
○ Ms. Camp wants to check for prerequisite knowledge in her students before she begins the unit on factoring.
○ Ms. Camp wants to measure growth in student achievement of these concepts, and scores on this test will serve as the students' knowledge baseline.

3.33 How confident are you in your answer to the above question?
○ Completely Unconfident
○ Mostly Unconfident
○ Neither Confident nor Unconfident

❍ Mostly Confident
❍ Completely Confident

3.34 To evaluate the effectiveness of the mathematics program for her gifted first graders, Ms. Allen gave them a standardized mathematics test normed for third graders. To decide how well her students performed, Ms. Allen compared her students' scores to those of the third-grade norm group. Why is this an incorrect application of standardized test norms?
❍ The norms are not reliable for first graders.
❍ The norms are not valid for first graders.
❍ Third grade mathematics items are too difficult for first graders.
❍ The time limits are too short for first graders.

3.35 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.36 When planning classroom instruction for a unit on arithmetic operations with fractions, which of these types of information have more potential to be helpful?
❍ norm-referenced information: describes each student's performance relative to other students in a group (e.g., percentile ranks, stanines), or
❍ criterion-referenced information: describes each student's performance in terms of status on specific learning outcomes (e.g., number of items correctly answered for each specific objective)
❍ Both types of information are equally useful in helping to plan for instruction.
❍ Neither, test information is not useful in helping to plan instruction.

3.37 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.38 Students' scores on standardized tests are sometimes inconsistent with their performances on classroom assessments (e.g., teacher tests or other in-class activities). Which of the following is not a reasonable explanation for such discrepancies?
❍ Some students freeze up on standardized tests, but they do fine on classroom assessments.
❍ Students often take standardized tests less seriously than they take classroom assessments.

○ Standardized tests measure only recall of information while classroom assessments measure more complex thinking.
○ Standardized tests may have less curriculum validity than classroom assessment.

3.39 How confident are you in your answer to the above question?
○ Completely Unconfident
○ Mostly Unconfident
○ Neither Confident nor Unconfident
○ Mostly Confident
○ Completely Confident

3.40 Elementary school teachers in the Baker School system collectively designed and developed new curricula in Reading, Mathematics, and Science that is based on locally developed objectives and objectives in state curriculum guides. The new curricula were not matched directly to the content of the fourth-grade standardized test. A newspaper reports the fourth-grade students in Baker Public Schools are among the lowest scoring districts in the State Assessment Program. Which of the following would invalidate the comparison between Baker Public Schools and other schools in the state?
○ The curriculum objectives of the other districts may more closely match those of the State Assessment.
○ Other school systems did not design their curriculum to be consistent with the State Assessment test.
○ Instruction in Baker schools is poor.
○ Other school systems have different promotion policies than Baker.

3.41 How confident are you in your answer to the above question?
○ Completely Unconfident
○ Mostly Unconfident
○ Neither Confident nor Unconfident
○ Mostly Confident
○ Completely Confident

3.42 Which of the following choices typically provides the most reliable student-performance information that a teacher might consider when assigning a unit grade?
○ Scores from a teacher-made test containing two or three essay questions related directly to instructional objectives of the unit.
○ Scores from a teacher-made 20 item multiple-choice test designed to measure the specific instructional objectives of the unit.
○ Oral responses to questions asked in class of each student over the course of the unit.
○ Daily grades designed to indicate the quality of in-class participation during regular instruction.

3.43 How confident are you in your answer to the above question?
○ Completely Unconfident

○ Mostly Unconfident
○ Neither Confident nor Unconfident
○ Mostly Confident
○ Completely Confident

3.44 A teacher gave three tests during a grading period and she wants to weight them all equally when assigning grades. The goal of the grading program is to rank order students on achievement. In order to achieve this goal, which of the following should be closest to equal?
○ Number of items.
○ Number of students taking each test.
○ Average scores.
○ Variation (range) of scores.

3.45 How confident are you in your answer to the above question?
○ Completely Unconfident
○ Mostly Unconfident
○ Neither Confident nor Unconfident
○ Mostly Confident
○ Completely Confident

3.46 When a parent asks a teacher to explain the basis for his or her child's grade, the teacher should:
○ explain that the grades are assigned fairly, based on the student's performance and other related factors.
○ ask the parents what they think should be the basis for the child's grade.
○ explain exactly how the grade was determined and show the parent samples of the student's work.
○ indicate that the grading scale is imposed by the school board and the teachers have no control over grades.

3.47 How confident are you in your answer to the above question?
○ Completely Unconfident
○ Mostly Unconfident
○ Neither Confident nor Unconfident
○ Mostly Confident
○ Completely Confident

3.48 Which of the following grading practices results in a grade that least reflects students' achievement?
○ Mr. Jones requires students to turn in homework; however, he only grades the odd numbered items.
○ Mrs. Brown uses weekly quizzes and three major examinations to assign final grades in her class.

- ○ Ms. Smith permits students to redo their assignments several times if they need more opportunities to meet her standards for grades.
- ○ Miss Engle deducts 5 points from a student's test grade for disruptive behavior.

3.49 How confident are you in your answer to the above question?

- ○ Completely Unconfident
- ○ Mostly Unconfident
- ○ Neither Confident nor Unconfident
- ○ Mostly Confident
- ○ Completely Confident

3.50 During the most recent grading period, Ms. Johnson graded no homework and gave only one end-of-unit test. Grades were assigned only on the basis of the test. Which of the following is the major criticism regarding how she assigned the grades?

- ○ The grades probably reflect a bias against minority students that exists in most tests.
- ○ Decisions like grade assignment should be based on more than one piece of information.
- ○ The test was too narrow in curriculum focus.
- ○ There is no significant criticism of this method providing the test covered the unit's content.

3.51 How confident are you in your answer to the above question?

- ○ Completely Unconfident
- ○ Mostly Unconfident
- ○ Neither Confident nor Unconfident
- ○ Mostly Confident
- ○ Completely Confident

3.52 In a routine conference with Mary's parents, Mrs. Estes observed that Mary's scores on the state assessment program's quantitative reasoning tests indicate Mary is performing better in mathematics concepts than in mathematics computation. This probably means that

- ○ Mary's score on the computation test was below average.
- ○ Mary is an excellent student in mathematics concepts.
- ○ the percentile bands for the mathematics concepts and computation tests do not overlap.
- ○ the mathematics concepts test is a more valid measure of Mary's quantitative reasoning ability.

3.53 How confident are you in your answer to the above question?

- ○ Completely Unconfident
- ○ Mostly Unconfident
- ○ Neither Confident nor Unconfident
- ○ Mostly Confident

❍ Completely Confident

3.54 Many states are revising their school accountability programs to help explain differences in test scores across school systems. Which of the following is not something that needs to be considered in such a program?
❍ The number of students in each school system.
❍ The average socio-economic status of the school systems.
❍ The race/ethnic distribution of students in each school system.
❍ The drop-out rate in each school systems.

3.55 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident
3.56 The following standardized test data are reported for John.
Subject -- Stanine Score
Vocabulary -- 7
Mathematics Computation -- 7
Social Studies – 7
Which of the following is a valid interpretation of this score report?
❍ John answered correctly the same number of items on each of the three tests.
❍ John's test scores are equivalent to a typical seventh grader's test performance.
❍ John had the same percentile rank on the three tests.
❍ John scored above average on each of the three tests.

3.57 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.58 Mr. Klein bases his students' grades mostly on graded homework and tests. Mr. Kaplan bases his students' grades mostly on his observation of the students during class. A major difference in these two assessment strategies for assigning grades can best be summarized as a difference in
❍ formal and informal assessment.
❍ performance and applied assessment.
❍ customized and tailored assessment.
❍ formative and summative assessment.

3.59 How confident are you in your answer to the above question?

❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.60 John scored at the 60th percentile on a mathematics concepts test and scored at the 57th percentile on a test of reading comprehension. If the percentile bands for each test are five percentile ranks wide, what should John's teacher do in light of these test results?
❍ Ignore this difference.
❍ Provide John with individual help in reading.
❍ Motivate John to read more extensively outside of school.
❍ Provide enrichment experiences for John in mathematics, his better performance area.

3.61 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.62 In some states testing companies are required to release items from prior versions of a test to anyone who requests them. Such requirements are known as
❍ open-testing mandates.
❍ gag rules.
❍ freedom-of-information acts.
❍ truth-in-testing laws.

3.63 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.64 Mrs. Brown wants to let her students know how they did on their test as quickly as possible. She tells her students that their scored tests will be on a chair outside of her room immediately after school. The students may come by and pick out their graded test from among the other tests for their class. What is wrong with Mrs. Brown's action?
❍ The students can see the other students' graded tests, making it a violation of the students' right of privacy.
❍ The students have to wait until after school, so the action is unfair to students who have to leave immediately after school.
❍ Mrs. Brown will have to rush to get the tests graded by the end of the school day,

hence, the action prevents her from using the test to identify students who need special help.
❍ The students who were absent will have an unfair advantage, because her action allows the possibility for these students to cheat.

3.65 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.66 A state uses its statewide testing program as a basis for distributing resources to school systems. To establish an equitable distribution plan, the criterion set by the State Board of Education provides additional resources to every school system with student achievement test scores above the state average. Which cliché best describes the likely outcome of this regulation?
❍ Every cloud has its silver lining.
❍ Into each life some rain must fall.
❍ The rich get richer and the poor get poorer.
❍ A bird in the hand is worth two in the bush.

3.67 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.68 In a school where teacher evaluations are based in part on their students' scores on a standardized test, several teachers noted that one of their students did not reach some vocabulary items on a standardized test. Which teacher's actions is considered ethical?
❍ Mr. Jackson darkened circles on the answer sheet at random. He assumed Fred, who was not a good student, would just guess at the answers, so this would be a fair way to obtain Fred's score on the test.
❍ Mr. Hoover filled in the answer sheet the way he thought Joan, who was not feeling well, would have answered based on Joan's typical in-class performance.
❍ Mr. Stover turned in the answer sheet as it was, even though he thought George, an average student, might have gotten a higher score had he finished the test.
❍ Mr. Lund read each question and darkened in the bubbles on the answer sheet that represented what he believed Felicia, a slightly below average student, would select as the correct answers.

3.69 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

3.70 Mrs. Overton was concerned that her students would not do well on the State Assessment Program to be administered in the Spring. She got a copy of the standardized test form that was going to be used. She did each of the following activities to help increase scores. Which activity was unethical?
❍ Instructed students in strategies on taking multiple choice tests, including how to use answer sheets.
❍ Gave students the items from an alternate form of the test.
❍ Planned instruction to focus on the concepts covered in the test.
❍ None of these actions are unethical.

3.71 How confident are you in your answer to the above question?
❍ Completely Unconfident
❍ Mostly Unconfident
❍ Neither Confident nor Unconfident
❍ Mostly Confident
❍ Completely Confident

APPENDIX B

CLASSROOM ASSESSMENT LITERACY INVENTORY (CALI) ITEMS AND THE *STANDARDS*

Appendix B
Classroom Assessment Literacy Inventory (CALI) Items and the *Standards*

1. Modified Classroom Assessment Li Teachers should be skilled in choosing assessment methods appropriate for instructional decisions.

   Items: 1, 2, 3, 4, 5

2. Teachers should be skilled in developing assessment methods appropriate for instructional decisions.

   Items: 6, 7, 8, 9, 10

3. The teacher should be skilled in administering, scoring and interpreting the results of both externally-produced and teacher-produced assessment methods.

   Items: 11, 12, 13, 14, 15

4. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum, and school improvement.

   Items: 16, 17, 18, 19, 20

5. Teachers should be skilled in developing valid pupil grading procedures which use pupil assessments.

   Items: 21, 22, 23, 24, 25

6. Teachers should be skilled in communicating assessment results to students, parents, other lay audiences, and other educators.

   Items: 26, 27, 28, 29, 30

7. Teachers should be skilled in recognizing unethical, illegal, and otherwise inappropriate assessment methods and uses of assessment information.

   Items: 31, 32, 33, 34, 35

APPENDIX C

Institutional Review Board (IRB) Forms

Appendix C
Institutional Review Board (IRB) Forms

Phase 1 Approval

IRB Level I, Category 2 Approval for Protocol Application #16-747

Protocol #16-747 - Entitled "An Examination of an Assessment Literacy Questionnaire"

We have assigned your application the following IRB number: **16-747**. Please reference this number when corresponding with our office regarding your application. The Kent State University Institutional Review Board has reviewed and approved your Application for Approval to Use Human Research Participants as Level I/Exempt from Annual review research. Your research project involves minimal risk to human subjects and meets the criteria for the following category of exemption under federal regulations:

- Exemption 2: Educational Tests, Surveys, Interviews, Public Behavior Observation. This application was approved on December 15, 2016. *Submission of annual review reports is not required for Level 1/Exempt projects. We do NOT stamp Level I protocol consent documents.*

For compliance with:

- DHHS regulations for the protection of human subjects (Title 45 part 46), subparts A, B, C, D & E

If any modifications are made in research design, methodology, or procedures that increase the risks to subjects or includes activities that do not fall within the approved exemption category, those modifications must be submitted to and approved by the IRB before implementation. Please contact an IRB discipline specific reviewer or the Office of Research Compliance to discuss the changes and whether a new application must be submitted. Visit our website for modification forms. Kent State University has a Federal Wide Assurance on file with the Office for Human Research Protections (OHRP); FWA Number 00001853.

To search for funding opportunities, please sign up for a free Pivot account at http://pivot.cos.com/funding_main. If you have any questions or concerns, please contact us at Researchcompliance@kent.edu or by phone at 330-672-2704 or 330.672.8058.

Doug Delahanty | IRB Chair |330.672.2395 | ddelahan@kent.edu
Tricia Sloan | Coordinator |330.672.2181 | psloan1@kent.edu
Kevin McCreary | Assistant Director | 330.672.8058 | kmccrea1@kent.edu
Paulette Washko | Director |330.672.2704| pwashko@kent.edu

Phase 1 Informed Consent

IRB # 16-747
Title: An Examination of an Assessment Literacy Questionnaire
Principle Investigator: Aryn C. Karpinski, Ph.D.
Co-Investigator: Kelli A. Ryan

You are invited to participate in a research study titled “Assessment Literacy Questionnaire.” This study is being conducted by Dr. Aryn C. Karpinski, who is an Assistant Professor in the School of Foundations, Leadership, and Administration (FLA) at Kent State University. This form provides you with information about the research project. It tells you what you will need to do to participate. It also tells you the associated risks and benefits of the research. Please read it carefully. It is important for you to fully understand the research in order to make an informed decision.

Purpose

We are doing this study to investigate what educators know about assessment. Educators vary widely in their knowledge of assessment, and we are trying to understand what concepts are understood and how confident educators are in their assessment knowledge.

Procedure

If you agree to participate, you will be asked to complete a multiple-choice questionnaire. Please set aside 30 minutes to complete this questionnaire. After you answer each question to the best of your ability, you will rate how confident you are in your answer. These data will help us understand what you know about assessment and how condiment you are in your assessment knowledge. All of your responses in this study are anonymous and will not be linked to your name or any other identifying information. The data will be stored along with randomly assigned ID numbers that are also not linked to or stored with your names or any other identifying information.

Privacy and Confidentiality

No information will be collected that allows us to connect your name to your data. You are also not asked to sign this informed consent document. Study data will be kept in password-protected folders on a password-protected computer that is kept under lock and key. Only the principle investigator, Dr. Aryn C. Karpinski, and the co-investigator, Kelli A. Ryan, will have access to these data.

Risks

There are no known risks associated with this study.

Benefits and Compensation

There are no direct benefits or compensation for participating in this study. While you will not experience any direct benefits from participation, information collected in this study may benefit others in the future by helping to better understand assessment literacy.

Voluntary Participation
Taking part in this research study is entirely up to you. You do not have to participate in the study, though. You can choose not to participate at all, or you can quit the study at any time if you want to. No matter what you decide, there will be no effect on your relationship with the researchers.

By completing this study, you are agreeing to participate in this research study. You may have a copy of this consent form, if you would like one.

Contact Information
If you have any questions about this research, please contact the principal investigator, Aryn C. Karpinski at akarpins@kent.edu. This project has been approved by the Kent State University Institutional Review Board. If you have any questions about your rights as a research participant or complaints about the research, you may call the IRB at 330-672-2704.

Aryn C. Karpinski, Ph.D.
Assistant Professor
Kent State University
Email: akarpins@kent.edu

Phase 1 Recruitment Email

Good Afternoon (name),

My name is Kelli A. Ryan and I am a Ph.D. student at Kent State University. I am currently investigating Assessment Literacy, which is how educators make sense of assessments to understand, interpret, and apply K-12 assessment for learning. This research is vital, as data-driven decisions and high-stakes assessments continue to grow. With your help, we can begin to identify areas in need of best practices for both in-service and pre-service educators. Please take 30 minutes to fill out this anonymous questionnaire (LINK TO ONLINE SURVEY).

Thank you for your time and participation!

Kelli A. Ryan

kryan19@kent.edu
PhD Student and Research Assistant
Evaluation and Measurement
Kent State University

Phase 1 Reminder Email

Good Afternoon (name),

Two weeks ago, you received an e-mail message asking you to assist us in research on Assessment Literacy, which is how educators make sense of assessments to understand, interpret, and apply K-12 assessment for learning. Participation requires filling out a 30-minute web-based questionnaire. If you have filled out the survey, thank you!

If you have not had a chance to take the survey yet, I would appreciate you're your time and completion the survey. This message has gone to everyone in the selected sample population. Since no personal data is retained with the surveys for reasons of confidentiality, we are unable to identify whether or not you have already completed the survey.

This research is vital, as data-driven decisions and high-stakes assessments continue to grow. With your help, we can begin to identify areas in need of best practices for both in-service and pre-service educators. Please take 30 minutes to fill out this anonymous questionnaire (LINK TO ONLINE SURVEY).

Thank you for your time and participation!

Kelli A. Ryan

kryan19@kent.edu
PhD Student and Research Assistant
Evaluation and Measurement
Kent State University

Phase 2 Approval

IRB Approval for Protocol #17-131
IRB # 17-131 entitled “An Investigation of Pre-Service Teacher Assessment Literacy and EdTPA Assessment Performance”

Hello,
I am pleased to inform you that the Kent State University Institutional Review Board reviewed and approved your Application for Approval to Use Human Research Participants as a Level II/Expedited, category 5 & 7 project. **Approval is effective for a twelve-month period:**

April 17th, 2017 through April 16th, 2018

For compliance with: DHHS regulations for the protection of human subjects (Title 45 part 46), subparts A, B, C, D & E

Federal regulations and Kent State University IRB policy require that research be reviewed at intervals appropriate to the degree of risk, but not less than once per year. The IRB has determined that this protocol requires an annual review and progress report. The IRB tries to send you annual review reminder notice by email as a courtesy. **However, please note that it is the responsibility of the principal investigator to be aware of the study expiration date and submit the required materials.** Please submit review materials (annual review form and copy of current consent form) one month prior to the expiration date. Visit our website for forms.

HHS regulations and Kent State University Institutional Review Board guidelines require that any changes in research methodology, protocol design, or principal investigator have the prior approval of the IRB before implementation and continuation of the protocol. The IRB must also be informed of any adverse events associated with the study. The IRB further requests a final report at the conclusion of the study. Kent State University has a Federal Wide Assurance on file with the Office for Human Research Protections (OHRP); FWA Number 00001853.

To search for funding opportunities, please sign up for a free Pivot account at http://pivot.cos.com/funding_main. If you have any questions or concerns, please contact the Office of Research Compliance at Researchcompliance@kent.edu or 330-672-2704 or 330-672-8058.

Bethany Holland | Assistant |330.672.2384| bhollan4_stu@kent.edu
Tricia Sloan | Coordinator |330.672.2181 | psloan1@kent.edu
Kevin McCreary | Assistant Director | 330.672.8058 | kmccrea1@kent.edu
Paulette Washko | Director |330.672.2704| pwashko@kent.edu
Doug Delahanty | IRB Chair |330.672.2395 | ddelahan@kent.edu

Phase 2 Informed Consent

IRB # ____________
Title: *An Investigation of Pre-Service Teacher Assessment Literacy and EdTPA Assessment Performance*
Principle Investigator: Aryn C. Karpinski (Kosmidis), Ph.D.
Co-Investigator: Kelli A. Ryan

You are invited to participate in a research study titled “An Investigation of Pre-Service Teacher Assessment Literacy and EdTPA Assessment Performance.” This study is being conducted by Dr. Aryn C. Karpinski, who is an Assistant Professor in the School of Foundations, Leadership, and Administration (FLA) at Kent State University, and Evaluation and Measurement Ph.D. Candidate Kelli Ryan. This form provides you with information about the research project for Kelli Ryan’s dissertation. It tells you what you will need to do to participate. It also tells you the associated risks and benefits of the research. Please read it carefully. It is important for you to fully understand the research in order to make an informed decision.

Purpose

We are doing this study to investigate what educators know about assessment. Educators vary widely in their knowledge of assessment, and we are trying to understand what concepts are understood and how confident educators are in their assessment knowledge.

Procedure

If you agree to participate, you will be asked to complete a multiple-choice questionnaire. Please set aside 30 minutes to complete this questionnaire. After you answer each question to the best of your ability, you will rate how confident you are in your answer. These data will help us understand what you know about assessment and how condiment you are in your assessment knowledge.

Upon completion of this questionnaire, your responses will be connected to your existing personal edTPA scores using your FlashLine ID. Once your questionnaire responses have been connected to your existing edTPA scores, any possible identifying information (i.e., your FlashLine ID) will be removed. All of your responses in this study will be made permanently anonymous at this point and will not be linked to your name or any other identifying information. The data will be stored along with randomly assigned ID numbers that are also not linked to or stored with your names or any other identifying information.

Privacy and Confidentiality

Upon completion of this survey and connection of your responses to existing edTPA scores, your FlashLine ID will be removed from your responses. Any information collected that allows us to connect your name to your data will be permanently removed. You are also not asked to sign this informed consent document. Study data will be kept in password-protected folders on a password-protected computer that is kept under lock and key. Only the principle investigator, Dr. Aryn C. Karpinski, and the co-investigator, Kelli A. Ryan, will have access to these data.

Risks
There are no known risks associated with this study.

Benefits and compensation
There are no direct benefits or compensation for participating in this study. While you will not experience any direct benefits from participation, information collected in this study may benefit others in the future by helping to better understand assessment literacy.

Voluntary Participation
Taking part in this research study is entirely up to you. You do not have to participate in the study, though. You can choose not to participate at all, or you can quit the study at any time if you want to. No matter what you decide, there will be no effect on your relationship with the researchers.

By completing this study, you are agreeing to participate in this research study. You may have a copy of this consent form, if you would like one.

Contact Information
If you have any questions about this research, please contact the principal investigator, Aryn C. Karpinski at akarpins@kent.edu or co-investigator, Kelli Ryan at kryan19@kent.edu. This project has been approved by the Kent State University Institutional Review Board. If you have any questions about your rights as a research participant or complaints about the research, you may call the IRB at 330-672-2704.

Aryn C. Karpinski, Ph.D.
Assistant Professor
Kent State University
Email: akarpins@kent.edu

Kelli Ryan
Ph.D. Candidate
Kent State University
Email: kryan19@kent.edu

Phase 2 Recruitment Email

Good Afternoon (name),

My name is Kelli A. Ryan and I am a Ph.D. Candidate at Kent State University. For my dissertation, I am currently investigating Assessment Literacy, which is how educators make sense of assessments to understand, interpret, and apply K-12 assessment for learning. This research is vital, as data-driven decisions and high-stakes assessments continue to grow. With your help, we can begin to identify areas in need of best practices for both in-service and pre-service educators. Please take 30 minutes to fill out this survey. You survey responses will be recorded using your FlashLine ID. Upon completion of the survey, your edTPA scores will be linked to your survey responses, de-identified, and made totally anonymous.

[INSERT LINK HERE]

Thank you for your time and participation!

Kelli A. Ryan

kryan19@kent.edu
PhD Candidate and Research Assistant
Evaluation and Measurement
Kent State University

Phase 2 Reminder Email

Good Afternoon (name),

Two weeks ago, you received an e-mail message asking you to assist us in research on Assessment Literacy, which is how educators make sense of assessments to understand, interpret, and apply K-12 assessment for learning. Participation requires filling out a 30-minute web-based survey. If you have filled out the survey, thank you!

If you have not had a chance to take the survey yet, I would appreciate you're your time and completion the survey. You survey responses will be recorded using your FlashLine ID. Upon completion of the survey, your edTPA scores will be linked to your survey responses, de-identified, and made totally anonymous.

This research is vital, as data-driven decisions and high-stakes assessments continue to grow. With your help, we can begin to identify areas in need of best practices for pre-service educators. Please take 30 minutes to fill out this survey.

[INSERT LINK HERE]

Thank you for your time and participation!

Kelli A. Ryan

kryan19@kent.edu
PhD Candidate and Research Assistant
Evaluation and Measurement
Kent State University

APPENDIX D

ITEM RE-NUMBERING AND PRINCIPAL COMPONENTS ANALYSIS (PCA) COMPONENTS

Appendix D
Item Re-Numbering and Principal Components Analysis (PCA) Components

1-35 Item Number (Pilot Study)	1-25 Item Number (2nd Phase Re-Numbered)	Component Number
1	1	2
2	2	2
3	3	2
4	4	2
5	5	2
6	X	
7	X	
8	6	1
9	7	2
10	8	2
11	9	1
12	10	1
13	11	2
14	X	
15	12	1
16	13	1
17	14	1
18	X	
19	X	
20	X	
21	X	
22	15	2
23	16	1
24	17	1
25	18	1
26	19	2
27	X	
28	X	
29	20	2
30	21	2
31	22	2
32	23	1
33	24	1
34	X	
35	25	1

APPENDIX E

ITEMS ACCORDING TO PRINCIPAL COMPONENTS ANALYSIS (PCA) COMPONENT LOADINGS

Appendix E
Items According to Principal Components Analysis (PCA) Component Loadings

PCA 1

A state uses its statewide testing program as a basis for distributing resources to school systems. To establish an equitable distribution plan, the criterion set by the State Board of Education provides additional resources to every school system with student achievement test scores above the state average. Which cliché best describes the likely outcome of this regulation?

Ms. Camp is starting a new semester with a factoring unit in her Algebra I class. Before beginning the unit, she gives her students a test on the commutative, associative, and distributive properties of addition and multiplication. Which of the following is the most likely reason she gives this test to her students?

Mrs. Brown wants to let her students know how they did on their test as quickly as possible. She tells her students that their scored tests will be on a chair outside of her room immediately after school. The students may come by and pick out their graded test from among the other tests for their class. What is wrong with Mrs. Brown's action?

During the most recent grading period, Ms. Johnson graded no homework and gave only one end-of-unit test. Grades were assigned only on the basis of the test. Which of the following is the major criticism regarding how she assigned the grades?

Students in Mr. Jakman's science class are required to develop a model of the solar system as part of their end-of-unit grade. Which scoring procedure below will maximize the objectivity of assessing these student projects?

When a parent asks a teacher to explain the basis for his or her child's grade, the teacher should:

When the directions indicate each section of a standardized test is timed separately, which of the following is acceptable test-taking behavior?

Ms. Gregory wants to assess her students' skills in organizing ideas rather than just repeating facts. Which words should she use in formulating essay exercises to achieve this goal?

Which of the following grading practices results in a grade that least reflects students' achievement?

that was going to be used. She did each of the following activities to help increase scores. Which activity was unethical?

Many teachers score classroom tests using a 100-point percent correct scale. In general, what does a student's score of 90 on such a scale mean?

To evaluate the effectiveness of the mathematics program for her gifted first graders, Ms. Allen gave them a standardized mathematics test normed for third graders. To decide how well her students performed, Ms. Allen compared her students' scores to those of the third-grade norm group. Why is this an incorrect application of standardized test norms?

PCA 2

What is the most effective use a teacher can make of an assessment that requires students to show their work (e.g., the way they arrived at a solution to a problem or the logic used to arrive at a conclusion)?

When scores from a standardized test are said to be “reliable,” what does it imply?

What is the most important consideration in choosing a method for assessing student achievement?

Several students in Ms. Atwell's class received low scores on her end-of-unit test covering multi-step story problems in mathematics. She wanted to know which students were having similar problems, so she could group them for instruction. Which assessment strategy would be best for her to use for grouping students?

Mrs. Bruce wished to assess her students' understanding of the method of problem solving she had been teaching. Which assessment strategy below would be most valid?

John scored at the 60th percentile on a mathematics concepts test and scored at the 57th percentile on a test of reading comprehension. If the percentile bands for each test are five percentile ranks wide, what should John's teacher do in light of these test results?

A teacher gave three tests during a grading period and she wants to weight them all equally when assigning grades. The goal of the grading program is to rank order students on achievement. In order to achieve this goal, which of the following should be closest to equal?

In a routine conference with Mary's parents, Mrs. Estes observed that Mary's scores on the state assessment program's quantitative reasoning tests indicate Mary is performing better in mathematics concepts than in mathematics computation. This probably means that

In some states testing companies are required to release items from prior versions of a test to anyone who requests them. Such requirements are known as

Ms. Green, the principal, was evaluating the teaching performance of Mr. Williams, the fourth-grade teacher. One of the things Ms. Green wanted to learn was if the students were being encouraged to use higher order thinking skills in the class. What documentation would be the most valid to help Ms. Green to make this decision?

Mr. Klein bases his students' grades mostly on graded homework and tests. Mr. Kaplan bases his students' grades mostly on his observation of the students during class. A major difference in these two assessment strategies for assigning grades can best be summarized as a difference in

Mr. Woodruff wanted his students to appreciate the literary works of Edgar Allen Poe. Which of his test items shown below will best measure his instructional goal?

At the close of the first month of school, Mrs. Friend gives her fifth grade students a test she developed in social studies. Her test is modeled after a standardized social studies test. It presents passages and then asks questions related to understanding and problem definition. When the test was scored, she noticed that two of her students—who had been performing well in their class assignments—scored much lower than other students. Which of the following types of additional information which would be most helpful in interpreting the results of this test?

REFERENCES

Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, Calif: Sage Publications.

American Federation of Teachers, National Council on Measurement in Education, National Education Association. (1990). *Standards for teacher competence in educational assessment of students*. Retrieved from http://buros.org/standards-teacher-competence-educational-assessment-students.

Anderson, J. R. (1982). Acquisition of cognitive skill. *Psychological Review*, *89*(4), 369.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*(4), 561-573.

Andrich, D. (1988). *Rasch models for measurement* (68). Sage.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care*, *42*(1), I-7.

Angelo, T. A., & Cross, K. P. (1991). *Classroom assessment techniques*. San Francisco, CA: Jossey-Bass.

Athanases, S. Z., Bennett, L. H., & Wahleithner, J. M. (2013). Fostering data literacy through preservice teacher inquiry in English language arts. *The Teacher Educator, 48*(1), 8-28.

Babakus, E., Ferguson Jr, C. E., & Jöreskog, K. G. (1987). The sensitivity of confirmatory maximum likelihood factor analysis to violations of measurement scale and distributional assumptions. *Journal of Marketing Research*, 222-228.

Ball, D. L., & McDiarmid, G. W. (1990). *The subject-matter preparation of teachers.* In W. R. Houston and M. H. J. Sikula (Eds.), Handbook of research on teacher education (pp. 437-449). New York, NY: Macmillan.

Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling*, *9*(1), 78-102.

Bandura. A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191-215.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: Freeman.

Bassiri, D. (2015). *Statistical properties of school value-added scores based on assessment of college readiness*. Iowa City, IA: ACT. Retrieved from: http://www.act.org/content/dam/act/unsecured/documents/ACT_RR2015-5.pdf

Betebenner, D. (2009). Norm and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42-51.

Betz, N. E., & Borgen, F. H. (2010). The CAPA integrative online system for college major exploration. *Journal of Career Assessment*, *18*(4), 317-327.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice*, *5*(1), 7-74.

Blank, R. K. (2010). State growth models for school accountability: Progress on development and reporting measures of student growth. *Council of Chief State School Officers*.

Bond, T., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences*. New York, NY: Routledge.

Boone, W. J., Staver J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. The Netherlands: Springer.

Boyles P. (2005). Assessment literacy. In Rosenbusch M. (Ed.), *National Assessment Summit Papers,* 11–15. Ames, IA: Iowa State University.

Bursal, M., & Paznokas, L. (2006). Mathematics anxiety and preservice elementary teachers' confidence to teach mathematics and science. *School Science and Mathematics*, *106*(4), 173-180.

Brentari, E., & Golia, S. (2007). Unidimensionality in the Rasch model: How to detect and interpret. *Statistica*, *67*(3), 253-261.

Brentari E, Golia S, Manisera M (2007) Models for categorical data: A comparison between the Rasch model and Nonlinear Principal Component Analysis. *Statistica & Applicazioni V,* (1), 53-77.

Briggs, S. R., & Cheek, J. M. (1988). On the nature of self-monitoring: Problems with assessment, problems with validity. *Journal of Personality and Social Psychology*, *54*(4), 663.

Bromme, R. (2001). Teacher expertise. In N. J. Smelser, & P. B. Baltes (Eds.), *International encyclopedia of the social and behavioral sciences,* 15(4),59-65. Amsterdam, Netherlands: Elsevier.

Brookhart, S. M., Loadman, W. E., & Miller, T. E. (1994). Relations between self-confidence and educational beliefs before and after teacher education. *College Student Journal*, *28*
(1), 57-66.

Brookhart, S. M. (1999). Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practice*, *18*(1), 5-13.

Brookhart, S. M. (2001). The "Standards" and classroom assessment research. *Educational Measurement: Issues and Practice*, *18*(1), 23-27.

Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, *30*(1), 3-12.

Browne, M. W. (1984). Asymptotically distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, *37*(1), 62-83.

Chmura Kraemer, H., Kiernan, M., Essex, M., & Kupfer, D. J. (2008). How and why criteria defining moderators and mediators differ between the Baron & Kenny and MacArthur approaches. *Health Psychology*, *27*(2), 101.

Coburn, C. E., Honig, M. I., & Stein, M. K. (2009). What's the evidence on districts' use of evidence? In Bransford, J. D., Stipek, D. J., Vye, N. J., Gomez, L. M., & Lam,

D. (Eds.), *The role of research in Educational Improvement,* 67-87. Cambridge, MA: Harvard Educational Press.

Coburn, C. E., & Turner, E. O. (2011). Research on data use: A framework and analysis. *Measurement*, *9*(4), 173-206.

Comrey, A. L., & Lee, H. B. (1992). Interpretation and application of Factor Analytic results. *A First Course on Factor Analysis, 2*, 250-254. Hillsdale, NJ: Lawrence Erlbaum.

Covay Minor, E., Desimone, L., Caines Lee, J., & Hochberg, E. D. (2016). Insights on how to shape teacher learning policy: The role of teacher content knowledge in explaining differential effects of professional development. *Education Policy Analysis Archives*, *24*.

Costello, A.B., & Osborne, J.W. (2005). Best practices in Exploratory Factor Analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(7), 1-9.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281.

D'Agostino, J. V., & Powers, S. J. (2009). Predicting teacher performance with test scores and grade point average: A meta-analysis. *American Educational Research Journal*, 46(1), 146-182.

Daniel, L. G., & King, D. A. (1998). Knowledge and use of testing and measurement literacy of elementary and secondary teachers. *The Journal of Educational Research*, *91*(6), 331-344.

Darling-Hammond, L., Chung, R., & Frelow, F. (2002). Variation in teacher preparation: How well do different pathways prepare teachers to teach? *Journal of Teacher Education*, *53*(4), 286-302.

DeAyala, R. J. (2013). *The theory and practice of item response theory*. New York, NY: Guilford Publications.

DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy, & Practice*, *17*(4), 419-438.

Dembo, M. H., & Gibson, S. (1985). Teachers' sense of efficacy: An important factor in school improvement. *The Elementary School Journal*, *86*(2), 173-184.

DePascale, C., Betebenner, D., Ryan, K., & Sharp, A. (2016). Building a conceptual framework for assessment literacy. Retrieved from: The National Center for the Improvement of Educational Assessment online website: www.nciea.org.

Dimitrov, D. M. (2013) *Quantitative research in education: Intermediate & advanced methods*. Oceanside, NY: Whittier Publications.

Druva, C. A., & Anderson, R. D. (1983). Science teacher characteristics by teacher behavior and by student outcome: A meta-analysis of research. *Journal of Research in Science Teaching*, *20*(5), 467-479.

Educational Testing Services. (2018). PRAXIS State Requirements Section. Retrieved from https://www.ets.org/praxis/states

edTPA, Pearson Education Inc. (2017). *The edTPA about section.* Retrieved from https://edTPA.com/.

Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford, England, U.K.: Oxford University Press.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272.

Fan, Y. C., Wang, T. H., & Wang, K. H. (2011). A web-based model for developing assessment literacy of secondary in-service teachers. *Computers & Education*, *57*(2), 1727-1740.

Frank, M., & Barzilai, A. (2004). Integrating alternative assessment in a project-based learning course for pre-service science and technology teachers. *Assessment & Evaluation in Higher Education*, *29*(1), 41-61.

Forero, C. G., Maydeu-Olivares, A., & Gallardo-Pujol, D. (2009). Factor Analysis with ordinal indicators: A Monte Carlo study comparing DWLS and ULS estimation. *Structural Equation Modeling*, *16*(4), 625-641.

Fox, R. J. (1983). *Confirmatory factor analysis*. Hoboken, NJ: John Wiley & Sons.

Garbett, D. (2003). Science education in early childhood teacher education: Putting forward a case to enhance student teachers' confidence and competence. *Research in Science Education, 33*(4), 467-481.

Gareis, C. R., & Grant, L. W. (2015). Assessment literacy for teacher candidates: A focused approach. *Teacher Educators' Journal*, 4-21.

Gerzon, N. (2015). Structuring professional learning to develop a culture of data use: Aligning knowledge from the field and research findings. *Teachers College Record*, *117*(4), 1-28.

Gibson, S., & Dembo, M. H. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology*, *76*(4), 569.

Greenberg, J., McKee, A., & Walsh, K. (2013). Teacher prep review: A review of the nation's teacher preparation programs. *National Council on Teacher Quality*.

Goren, P. (2012). Data, data, and more data – What's an educator to do? *American Journal of Education*, *118*(2), 233-237.

Gotch, C. M., & French, B. F. (2014). A systematic review of assessment literacy measures. *Educational Measurement: Issues and Practice, 33*(2), 14-18.

Graham, R. C., Burgoyne, N., Cantrell, P., Smith, L., St Clair, L., & Harris, R. (2009). Measuring the TPACK confidence of inservice science teachers. *TechTrends*, *53*(5), 70-79.

Grimby, G., Tennant, A., & Tesio, L. (2012). The use of raw scores from ordinal scales: time to end malpractice? *Journal of Rehabilitation Medicine*, 44(2), 97-98.

Guadagnoli, E., & Velicer, W. F. (1988). Relation to sample size to the stability of component patterns. *Psychological Bulletin*, *103*(2), 265.

Gummer, E. S., & Mandinach, E. B. (2015). Building a conceptual framework for data literacy. *Teachers College Record, 117*(4), 1-22.

Hahs-Vaughn, D. L., & Lomax, R. G. (2013). *An introduction to statistical concepts*. New York, NY: Routledge.

Hayes, A. F. (2012). PROCESS: A versatile computational tool for observed variable mediation, moderation, and conditional process modeling. Retrieved from http://www.afhayes.com/ public/process2012.pdf

Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement*, *66*(3), 393-416.

Hills, J. R. (1991). Apathy concerning grading and testing. *Phi Delta Kappan*, *72*(7), 540-45.

Hooper, D., Coughlan, J., & Mullen, M. (2008). Structural equation modelling: Guidelines for determining model fit. *Journal of Business Research Methods*, 6(1).

Horn, I. S., & Little, J. W. (2010). Attending to problems of practice: Routines and resources for professional learning in teachers' workplace interactions. *American Educational Research Journal*, *47*(1), 181-217.

Hoy, A. W., & Spero, R. B. (2005). Changes in teacher efficacy during the early years of teaching: A comparison of four measures. *Teaching and Teacher Education*, 21(4), 343-356.

Hu, L. T., Bentler, P. M., & Kano, Y. (1992). Can test statistics in covariance structure analysis be trusted? *Psychological Bulletin*, *112*(2), 351.

Inbar-Lourie, O. (2008). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, *25*(3), 385-402.

James, W. (1985). *Psychology: The briefer course*. Notre Dame, IN: University of Notre Dame Press.

Jimerson, J. B., & Wayman, J. C. (2015). Professional learning for using data: Examining teacher needs and supports. *Teachers College Record, 117*(4), 1-36.

Joreskog, K. G. (1969). Statistical analysis of sets of congeneric tests. *ETS Research Report Series, 2*.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8: User's reference guide*. Scientific Software International.

Jöreskog, K. G. (1999). How large can a standardized coefficient be (Unpublished technical report). Retrieved from: http://www.ssicentral.com/lisrel/techdocs/HowLargeCana StandardizedCoefficientbe.

Jung, S. (2013). Exploratory factor analysis with small sample sizes: A comparison of three approaches. *Behavioral Processes*, 97, 90-95.

Kahl, S. R., Hofman, P., & Bryant, S. (2013). Assessment literacy standards and performance measures for teacher candidates and practicing teachers . Dover, NH: Measured Progress

Kazdin, A.E. (2003). *Research design in clinical psychology*. Boston, MA: Allyn & Bacon.

Keith, T. Z. (2014). *Multiple regression and beyond: An introduction to multiple regression and structural equation modeling*. New York, NY: Routledge.

Kim, H. J. (2008). Common factor analysis versus principal component analysis: choice for symptom cluster research. *Asian Nursing Research*, *2*(1), 17-24.

Kline, R. B. (2016). *Principles and practice of structural equation modeling*, *4*. New York: The Guilford Press.

Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education*, *22*(3), 255-276.

Kuder, G. F., & M. W. Richardson. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*, 151-160.

Lachat, M. A., & Smith, S. (2005). Practices that support data use in urban high schools. *Journal of Education for Students Placed at Risk*, *10*(3), 333-349.

Lambert, J. D. (1991). *Numerical methods for ordinary differential systems: The initial value problem*. Hoboken, NJ: John Wiley & Sons.

Lawrence, I. M., & Dorans, N. J. (1987*). An assessment of the dimensionality of SAT-Mathematical.*

Leahy S., Lyon C., Thompson M., Wiliam D. (2005). Classroom assessment minute by minute, day by day. *Educational Leadership*, 63, 18–24

Ledwell, K., & Oyler, C. (2016). Unstandardized responses to a “standardized” test: The edTPA as gatekeeper and curriculum change agent. *Journal of Teacher Education*, *67*(2), 120-134.

Li, C. H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods*, *48*(3), 936-949.

Linacre, J. M. (1998). Structure in Rasch residuals: Why principal components analysis? *Rasch Measurement Transactions*, *12*(2), 636.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement*, *2*, 266-283.

Linacre J. M., Wright B. D. (2000). *A user's guide to Winsteps Rasch-model computer programs* be (Unpublished technical report). Chicago, IL: MESA Press.

Linacre, J. M. (2012). *Winsteps Rasch Tutorial 2* (Unpublished technical report). Retrieved from: http://www.winsteps.com/tutorials.html.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling*, *9*(2), 151-173.

Little, T. D., Rhemtulla, M., Gibson, K., & Schoemann, A. M. (2013). Why the items versus parcels controversy needn’t be one. *Psychological Methods*, *18*(3), 285.

Love, N. (2004). Taking data to new depths. *Journal of Staff Development*, *25*(4), 22-26.

Lukin, L. E., Bandalos, D. L., Eckhout, T. J., & Mickelson, K. (2004). Facilitating the development of assessment literacy. *Educational Measurement: Issues and Practice*, *23*(2), 26-32.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84.

Main, S., & Hammond, L. (2008). Best practice or most practiced? Pre-service teachers' beliefs about effective behaviour management strategies and reported self-efficacy. *Australian Journal of Teacher Education*, *33*(4).

Mandinach, E. B., Gummer, E. S., & Friedman, J. M. (2015). How can schools of education help to build educators' capacity to use data: A systemic view of the issue. *Teachers College Record*, *117*(4), 1-50.

Marsh, H. W. (1993). Academic self-concept: theory, measurement, and research. In J. Suls (Ed.), *Psychological perspectives on the self*. 4, 59-98. Hillsdale, NJ: Lawrence Erlbaum.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, *103*(3), 391.

Marsh, J.A., Pane, J.F., & Hamilton, L.S. (2006). *Making sense of data-driven decision making in education: Evidence from Recent RAND research.* Santa Monica, CA: RAND Corporation.

Maslow, A. H. (1954). The instinctoid nature of basic needs. *Journal of Personality*, *22*(3), 326-347.

Matsunaga, M. (2010). How to Factor-Analyze your data right: Do's, don'ts, and how-to's. *International Journal of Psychological Research*, *3*(1).

Maurer, T. J., & Pierce, H. R. (1998). A comparison of Likert scale and traditional measures of self-efficacy. *Journal of Applied Psychology*, *83*(2), 324.

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating Value-Added Models for Teacher Accountability*. RAND Corporation : Santa Monica, CA.

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, *23*(1), 1-21.

McMillan, J. H. (2000). Fundamental assessment principles for teachers and school administrators. *Practical Assessment, Research, & Evaluation*, *7*(8).

McMillan, J. H. (2003). Understanding and improving teachers' classroom assessment decision making: Implications for theory and practice. *Educational Measurement: Issues and Practice*, *22*(4), 34-43.

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation.* John Wiley & Sons.

Mertler, C.A. (1998). Classroom assessment practices of Ohio teachers. Proceedings from*: Mid-Western Educational Research Association*,.Chicago, IL.

Mertler, C. A. (1999). Assessing student performance: A descriptive study of the classroom assessment practices of Ohio teachers. *Education*, *120*(2), 285-285.

Mertler, C. A. (2003). Preservice versus inservice teachers' assessment literacy: Does classroom experience make a difference? Proceedings from: *Mid-Western Educational Research Association*, Columbus, OH.

Mertler, C. A. (2004). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education*, 49-64.

Mertler, C.A. & Campbell, C.S. (2005). Measuring teachers' knowledge and application of classroom assessment concepts: Development of the Assessment Literacy Inventory. Proceeding from *American Educational Research Association*, Montreal, Quebec, Canada.

Michigan Assessment Consortium. (2015). Assessment literacy standards for teachers. The 2010 annual report on teacher education. *National Council for Accreditation of Teacher Education.*

Muthen, B. (1983). Latent variable structural equation modeling with categorical data. *Journal of Econometrics*, *22*(1-2), 43-65.

Muthen, B. (1983b). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115-132.

National Council of Teacher Quality (2013). The 2013 annual teacher preparation review: A review of the nation’s teacher preparation programs. Retrieved from http://www.nctq.org/reports.do.

Newton, S. (2010). Preservice performance assessment and teacher early career effectiveness: Preliminary findings on the performance assessment for California teachers. *Stanford, CA: Stanford University, Stanford Center for Assessment, Learning, and Equity.*

No Child Left Behind Act of 2001, P.L. 107-110, 20 U.S.C. § 6319 (2002).

Northwest Evaluation Association (2016). Make assessment work for all students: Multiple measures matter. Washington, DC: *Gallup*.

Nunnally, J. (1978). *Psychometric methods*. New York, NY: McGraw-Hill.

Obenchain, R. L. (1975). Residual optimality: Ordinary vs. weighted vs. biased least squares. *Journal of the American Statistical Association*, *70*(350), 375-379.

Ohio Department of Education. (2015). *URM Modeling Approach for Value-Added*. SAS EVASS. Retrieved from: https://education.ohio.gov/getattachment/Topics/Data/Report-Card-Resources/Ohio-Report-Cards/Value-Added-Technical-Reports-1/URM-Modeling-Approach.pdf.aspx

Ohio Department of Education. (2017). Teaching section. Retrieved from https://education.ohio.gov/.

Olsson, U. (1979). Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, *44*(4), 443-460.

Orcan, F. (2013). *Use of item parceling in structural equation modeling with missing data* (Doctoral dissertation). Retrieved from: The Florida State University.

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*, 543-578.

Park, V., & Datnow, A. (2009). Co-constructing distributed leadership: District and school connections in data-driven decision-making. *School Leadership and Management*, *29*(5), 477-494.

Pearson, K. (1900). Mathematical contributions to the theory of evolution VII on the correlation of characters not quantitatively measureable. *Philosophical Transactions of the Royal Society*, *195*, 1-47.

Pecheone, R. L., & Whittaker, A. (2016). Well-prepared teachers inspire student learning. *Phi Delta Kappan*, *97*(7), 8-13.

Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher*, *6*(1), 21-27.

Popham, W. J. (2003). *Test better, teach better: The instructional role of assessment*. Alexandria, VA: Association for Supervision and Curriculum and Development.

Popham, W. J. (2004). Why assessment illiteracy is professional suicide. *Educational Leadership*, *62*(1), 82.

Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory into Practice*, *48*(1), 4-11.

Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator*, *46*(4), 265-273.

Popham, W. J. (2013). *Classroom assessment: What teachers need to know*. Needham Heights, MA: Pearson.

Reynolds, C. R., Livingston, R. B., Willson, V. L., & Willson, V. (2010). *Measurement and Assessment in Education.* Princeton, NJ: Pearson.

Rigdon, E. E., & Ferguson Jr, C. E. (1991). The performance of the polychoric correlation coefficient and selected fitting functions in Confirmatory Factor Analysis with ordinal data. *Journal of Marketing Research*, 491-497.

Robinson, E. S. (2017). *Science Content Knowledge: A Component of Teacher Effectiveness in a Primary School in Jamaica* (Doctoral Dissertation). Walden University.

Rogosa, D. (1981). On the relationship between the Johnson-Neyman region of significance and statistical tests of parallel within-group regressions. *Educational and Psychological Measurement*, *41*(1), 73-84.

Sander, P., & Sanders, L. (2003). Measuring confidence in academic study: A summary report. *Electronic Journal of Research in Educational Psychology and Psychopedagogy*, *1*(1), 1-17.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In Alexander von Eye and Clifford C. Clogg (Eds.), *Latent Variables Analysis: Application to Developmental Research*, 399-419. Thousand Oaks, CA: SAGE Publications.

Savalei, V., Bonett, D. G., & Bentler, P. M. (2013). CFA with binary variables in small samples: A comparison of two methods. *Frontiers in Psychology*, 5, 1515

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting Structural Equation Modeling and Confirmatory Factor Analysis results: A review. *The Journal of Educational Research*, *99*(6), 323-338.

Schunk, D. H. (1981). Modeling and attributional effects on children's achievement: A self-efficacy analysis. *Journal of Educational Psychology*, *73*, 93-105.

Schunk, D. H. (1996). *Self-efficacy for learning and performance.* Presentation. American Educational Research Association, New York, NY.

Schunk, D. H., & Hanson, A. R. (1985). Peer models: Influence on children's self-efficacy and achievement. *Journal of Educational Psychology*, *77*(3), 313.

Schunk, D. H., Hanson, A. R., & Cox, P. D. (1987). Peer-model attributes and children's achievement behaviors. *Journal of Educational Psychology*, *79*(1), 54.

Schumacker, R. E., & Lomax, R. G. (2012). *A beginner's guide to Structural Equation Modeling*. New York, NY: Routledge Academic.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3/4), 591-611.

Siegel, M. A., & Wissehr, C. (2011). Preparing for the plunge: Preservice teachers' Assessment Literacy. *Journal of Science Teacher Education*, *22*(4), 371-391.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453.

Smith, R.M., Linacre, J.M., and Smith, Jr., E.V. (2003). Guidelines for Manuscripts. *Journal of Applied Measurement, 4*, 198-204.

Smith, C. D., Worsfold, K., Davies, L., Fisher, R., & McPhail, R. (2013). Assessment literacy and student learning: the case for explicitly developing students 'assessment literacy'. *Assessment & Evaluation in Higher Education*, 38(1), 44-60.

Stanford Center for Assessment, Learning, and Equity. (2017). About section. Retrieved from https://scale.stanford.edu/.

Stanford Center for Assessment, Learning and Equity. (2013). 2013 edTPA field test: Summary report. Stanford, CA: *Stanford Center for Assessment, Learning and Equity.*

Stevens, J. P. (2009). *Applied Multivariate Statistics for the social sciences*. New York, NY: Routledge.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan*, *72*(7), 534-39.

Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, *18*(1), 23-27.

Supovitz, J. A., & Klein, V. (2003). Mapping a course for improved student learning: How innovative schools systematically use student performance data to guide improvement. Philadelphia, PA: *Consortium for Policy Research in Education.*

Supovitz, J. (2010). Knowledge-based organizational learning for instructional improvement. In *Second international handbook of educational change,* 701-723. Netherlands: Springer.

Suskie, L. (2009). Using assessment results to inform teaching practice and promote lasting learning. In *Assessment, Learning and Judgement in Higher Education.* 1-20. Springer, Dordrecht.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment for learning. *Phi Delta Kappan*, *83*(10), 758.

Tabachnick, B.G., & Fidell, L.S. (2001). *Using Multivariate Statistics*, *4*. Needham Heights, MA: Allyn & Bacon.

Taras, M. (2005). Assessment–summative and formative–some theoretical reflections. *British Journal of Educational Studies*, *53*(4), 466-478.

Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, *29*, 21-36.

Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. American Psychological Association.

Tschannen-Moran, M., & Hoy, A. W. (2001). Teacher efficacy: Capturing an elusive construct. *Teaching and Teacher Education*, *17*(7), 783-805.

Volante, L., & Fazio, X. (2007). Exploring teacher candidates' assessment literacy: Implication for teacher education reform and professional development. *Canadian Journal of Education*, *30*(3), 749.

Walsh, E., & Isenberg, E. (2015). How does value added compare to student growth percentiles? *Statistics and Public Policy*, *2*(1), 1-13.

Wang, T. H., Wang, K. H., & Huang, S. C. (2008). Designing a web-based assessment environment for improving pre-service teacher assessment literacy. *Computers & Education*, *51*(1), 448-462.

Watson, J. M. (2001). Profiling teachers' competence and confidence to teach particular mathematics topics: The case of chance and data. *Journal of Mathematics Teacher Education*, *4*(4), 305-337.

Whitt, C., & Abigail, K. (2016). *A Structural Model of Elementary Teachers' Knowledge, Beliefs, and Practices for Next Generation Science Teaching* (Doctoral Dissertation). Wright State University.

Wiliam, D., & Black, P. (1996). Meanings and consequences: A basis for distinguishing formative and summative functions of assessment? *British Educational Research Journal*, 22, 537-548.

Wiliam, D., & Leahy, S. (2007). A theoretical foundation for formative assessment. In J. H. McMillan (Ed.), *Formative classroom assessment: Research, theory and practice*. New York, NY: Teachers College Press.

Wolters, C. A., & Daugherty, S. G. (2007). Goal structures and teachers' sense of efficacy: Their relation and association to teaching experience and academic level. *Journal of Educational Psychology*, *99*(1), 181.

Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions*, *10*(3), 509-511.

Wright, B. D., Linacre, J. M., Gustafson, J. E., & Martin-Lof, P. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Wright, B. D., & Stone, M. H. (1979). *Best test design: Rasch measurement*. Mesa Press. Chicago, IL: Mesa Press.

Xu, Y., & Brown, G. T. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*, 149-162.

Yang-Wallentin, F., Jöreskog, K. G., & Luo, H. (2010). Confirmatory Factor Analysis of ordinal variables with misspecified models. *Structural Equation Modeling*, *17*(3), 392-423.

Zhang, Z. (1996). Teacher assessment competency: A Rasch model analysis. Proceedings from *American Educational Research Association.*, New York, NY.

Zhang, Z., & Burry-Stock, J. A. (1997). Assessment practices inventory: A multivariate analysis of teachers' perceived assessment competency. Proceedings from *the National Council on Measurement in Education.* Chicago, IL.

Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education*, *16*(4), 323-342.